The international journal of science / 28 May 2020

nature

index
South Korea

HUMAN RESOURCE

gnomAD: a catalogue of variation in the human genome



Coronavirus
The drive to combat misinformation about COVID-19

Photocatalytic power
Hydrogen split from water with near-perfect efficiency

Scent assessment
Sniff test helps gauge level of consciousness after brain injury

Vol. 581, No. 7809
nature.com

The international journal of science / 28 May 2020

nature

Coronavirus misinformation needs engagement

Researchers must be open and transparent – acknowledge what is known and what isn't.

The past few weeks have seen an explosion in misleading claims about COVID-19. These are mostly online, and many are intended to sow doubts about vaccination as a way to protect against infection. For the individuals and organizations involved in such disinformation, the pandemic is a gilded opportunity. They are capitalizing on both the many unknowns about the SARS-CoV-2 virus and the disease it causes, as well as the many legitimate questions about safety and efficacy as vaccines are being developed at unprecedented speed.

Vaccines must be safe and effective. Once (and only once) this is proven, immunization campaigns need to be comprehensive to succeed. But this presents many challenges. For low-income countries, and in those without universal health care, a key obstacle is ensuring that vaccines are available and affordable. For certain higher-income countries – for example some in Europe – the challenge for coronavirus will be to overcome scepticism about vaccines, which is being fuelled by false information.

Researchers can play a part. Knowing what to do in the middle of a pandemic isn't straightforward. But for those considering how to respond to the kinds of questions that everyone is asking, and what to do about disinformation, there are ways to help.

Tackling disinformation

As *Nature* reports on page 371, misinformation (false information) and disinformation (information that is deliberately misleading) are complex. Some politicians are spreading virus disinformation to burnish their image and influence among their supporters. There are organizations that have set up disinformation websites – including money-making scams. Very little, if any, of this information will have been put through an open process of verification and review. For consumers, it can be a double whammy – they are paying, and also being misinformed or misled.

Public-health agencies and technology firms are aware of the harm being done and are working to respond. To their credit, platforms such as Facebook and YouTube are more active in taking down posts where there is a clear risk to public health. When questions such as "are vaccines safe" are typed into Google, the search algorithms are listing sources that provide evidence-based information. But for every item of misinformation and disinformation that are dealt with, more pop up. Moreover, sites have discovered ways to circumvent artificial-intelligence tools and harried moderators, and that makes the role of human fact-checkers more important.

> "Involving communities helps to create and maintain trust."

One thing that researchers can do is to work with organizations that are responding to disinformation. They can support or join in the work of professional fact-checkers, journalists and academics, doggedly following bots and disinformation-news sites, flagging their content to the media organizations and social-media firms that host these sites. Groups all over the world are involved in this response – including professional bodies, learned societies and media-facing organizations. The work they do is labour-intensive and can seem never-ending, but it is needed now more than ever.

Public engagement and transparency

Many people are asking important questions on subjects such as the safety of proposed vaccines, the security of contact-tracing apps and how intellectual property rights and profits from new drugs and vaccines will be shared. These are questions that researchers from fields such as public health, data security and health-care finance are also asking. If they are not already doing so, now is the time for these and other researchers to expand their public engagement.

It might be that a definite answer isn't known, or that there are a range of possible answers. That is often the case in science. The study and practice of public engagement in science has shown that involving communities in the kinds of conversations that researchers have – conversations about how scientists search for evidence, and being transparent about what is known and not known – all helps to create and maintain trust.

A year ago, the UK biomedical funding charity Wellcome published the results of a large global survey into vaccines, involving 140,000 participants in 140 countries. It found that around 80% of respondents considered vaccines safe and effective. Confidence was highest in low-income countries – notably Bangladesh and Rwanda – where public-awareness campaigns against infectious diseases such as malaria, typhoid and hepatitis are common.

By contrast, confidence in the importance of vaccines was lower in Europe, where populations are comparatively free of infectious diseases, but now have some of the highest deaths and infections from COVID-19. Some 22% of respondents from Europe are not confident that vaccines are safe, and this figure increases to 33% for France. Wellcome's findings reflect those from the European Commission's own *State of Vaccine Confidence* report from 2018. Across the European Union, health ministries are unable to meet their own target – set after the 2009 H1N1 swine flu outbreak – of vaccinating 75% of over-65s against flu.

Last November, Heidi Larson, an anthropologist at the London School of Hygiene and Tropical Medicine – and a co-author of the European Commission report – warned in an interview with *Nature* that if there is "another very serious influenza pandemic sooner or later, and if the public opt to forgo vaccination the way they did during the 2009

© 2020 Springer Nature Limited. All rights reserved.

nature

swine-flu pandemic, we're in deep trouble" (S. el-Showk *Nature* **575**, S57; 2019).

That is why there is work to be done. When it comes to communicating emerging information on research, the lessons from studies and from past practice are clear: not to over-promise, nor oversell and to emphasize what is known and what isn't. In the case of vaccines, it means being as transparent as possible about how vaccines are made, how they work, what they contain and how they will be tested, and always being upfront about the evidence for their effectiveness, possible risks and side effects.

Researchers should play a part — no matter how small — in the response to misinformation and disinformation. We need to build a society that is resilient to falsehoods about COVID-19, a task that will only become more vital as vaccines near.

Milestone in human genetics highlights diversity gap

Landmark study identifies genes that it seems people can and cannot live without. But such data still need to be more representative.

From the time that the nineteenth-century monk Gregor Mendel squinted at the pea plants in his garden and wondered why some had white flowers or wrinkled seeds, it has been a tradition in biology to observe what goes awry when a DNA sequence is altered — whether that variation occurs naturally or through human intervention.

Although geneticists have long been able to introduce genetic mutations into model organisms such as the fruit fly — first with X-rays or chemicals, and now with more sophisticated gene-editing tools — where humans are concerned, the toolbox is more limited. Researchers clearly cannot intentionally introduce mutations into humans; instead, they must use what nature provides. As a result, they comb through genomes in search of variations in DNA sequences, and use statistical tools to determine whether those variations contribute to traits and diseases. As genome sequencing has become quicker and cheaper, those studies have become bigger and more complex.

This week, three journals in the *Nature* family are publishing the results of the latest effort: a study of a staggering 125,748 exomes (the part of the genome that codes for proteins) and 15,708 whole genomes (see go.nature.com/2zgfxr2). The study — the most extensive publicly accessible analysis carried out so far — sheds light on which genes are essential and which a person might be able to live without. The results are compiled in the Genome Aggregation Database (gnomAD) and will help researchers to better understand the roots of genetic disorders, and, eventually, how best to treat them. That mutations can inactivate genes is hardly new, but this study adds to the surprisingly long list of mutations that can obliterate a gene's function without causing obvious harm. The study also identified a flurry of genes that are probably vital for life, because people rarely harbour drastic mutations predicted to cause 'loss of function' in these genes.

> **"Around half of the samples were donated by people of European descent."**

The study's large scale made it possible for the authors to devise a measure of how tolerant to loss-of-function mutations a given gene might be. This is a useful tool with which to study the function of known and newly identified genes, to pinpoint candidate disease-causing mutations, and to find new drug targets in the human genome.

One example is the team's evaluation of the gene *LRRK2*, which has been implicated in Parkinson's disease (N. Whiffin *et al. Nature Med.* https://doi.org/10.1038/s41591-020-0893-5; 2020). DNA variants that increase the activity of the LRRK2 protein have been associated with a higher risk of the disease, leading scientists to think that a drug that switches the gene off could be beneficial. But would turning off *LRRK2*, which is active in the brain, as well as in other tissues, be dangerous? Looking through gnomAD's 140,000 genomes and exomes, the authors found many naturally occurring DNA sequence variants that switch off *LRRK2*. That suggests — at least in principle — that a drug that can mimic this effect might not be harmful.

To answer such questions, a very large number of samples is needed, in part because DNA sequence variations that wipe out the function of an important gene are likely to be rare. This means that the more genomes scientists can analyse, the more variants they can find and the better they can pick apart the effects of each one. But such projects also need a greater diversity of participants than they have had thus far.

In the current studies, around half of the samples were donated by people of European descent. Although this is an improvement on previous studies, people from regions such as Central Asia, Oceania, the Middle East and much of Africa are almost absent. This means researchers are probably missing variants that are important for understanding gene function — and disease risk — in these regions. This is something that consortium members recognize, but progress is slow. Researchers and funders must incentivize such work to ensure that it continues to expand.

The gnomAD database is an outstanding resource. The willingness of participants to contribute — along with the willingness of researchers to share — has been key to its success. Further insights will come from combining sequence data with clinical information. Projects such as the Estonian Biobank, which includes more than 200,000 participants, and the UK Biobank, which has DNA and health information from 500,000 people, are paving the way. But such efforts need the involvement of more-diverse populations.

With these improvements, researchers will be able to maximize the contribution of everyone who provided their DNA samples to improve our knowledge of human biology and to fully harness genetic differences to benefit us all.

© 2020 Springer Nature Limited. All rights reserved.

World view



By Andy Haldane

Model lives and livelihoods in lockstep

Economists must improve tools to weigh trade-offs between health and wealth.

Last month I lost a co-author and friend whose interdisciplinary work now seems chillingly prescient. Robert May, once chief scientific adviser to the UK government and president of the Royal Society, did ground-breaking work on disease contagion, among other things.

A decade ago, May and I published a paper in *Nature* using models drawn from epidemiology to understand the dynamics of the global financial crisis of 2008–09, and appropriate policy responses to it. This instigated new thinking on the modelling of economies and financial systems. It also helped to demonstrate the value of cross-disciplinary research involving economics. Neither had featured strongly before the crisis.

The work made clear that economies and financial sectors, as connected social systems, have classic robust-yet-fragile properties (A. G. Haldane and R. M. May *Nature* **469**, 351–355; 2011). They are prone to periodic, self-generated tipping points, in which they shift quickly into a new state from which they can't return. Avoiding these means strengthening the systems' resilience by building up buffers to cushion stress, particularly among the largest, most connected financial firms – super-spreaders, if you will.

Those findings were the opposite of the orthodoxy pre-crisis, when super-spreader banks, and the financial system as a whole, ran with dangerously slender buffers of capital and liquidity. Encouragingly, in the decade since we published our paper, financial policy has been radically reformed and significantly larger buffers have built up.

It is just as well. The COVID-19 crisis has again exposed the fragilities of social and economic systems and how they can operate on a knife edge. This time, the source of the threat is public health rather than financial wealth, but again the risk is systemic and chronic. Public-health concerns have, rightly, taken priority in the setting of policy.

But the social-distancing measures put in place around the world to limit viral transmission have come at a significant economic and financial cost. As businesses and households have locked down, there has been a collapse in global activity and spending, unprecedented in its speed and severity. This has prompted a similarly unprecedented loosening of fiscal and monetary policies.

Public policy remains on the horns of a dilemma in many countries. It is navigating a narrow path that can seem to pit livelihoods against lives, the needs of older generations against those of younger ones, the health benefits of physical distance against the social benefits of societal cohesion. These are trade-offs the like and extent of which policymakers have never seen.

> "The COVID-19 crisis has exposed the fragilities of social and economic systems."

Andy Haldane is chief economist of the Bank of England in London.
e-mail: andy.haldane@bankofengland.co.uk

It is here that economists have an important role: we have always articulated and calibrated such trade-offs. This has given economists influence over some of the most difficult choices facing policymakers in recent decades, including how to tackle the other existential crisis – climate change.

Encouragingly, economists have quickly risen to the challenge. As one example, the Centre for Economic Policy Research in London is now publishing articles several times each week, gathering together economic research on the financial and social impact of COVID-19.

These insights have often come from combining modelling approaches from the natural and social sciences. For example, embedding an SIR (Susceptible, Infected, Recovered) model of disease dynamics in a general equilibrium model of people's spending decisions allows us to capture and calibrate some of the difficult trade-offs.

Despite this rapid progress, these models are still too fledgling and crude to provide robust advice to policymakers weighing economic and health outcomes – for example, deciding when and how best to ease social distancing. One of the most pressing analytical challenges ahead, then, is to advance these models.

After the 2008–09 financial crisis, reform focused on protecting those financial-sector activities that were most crucial to the public. These were not the high-risk, high-return activities that sowed the seeds of the crash. They were the everyday essentials of banking – the making of payments and loans to individuals and businesses. Post-crisis, these activities were ring-fenced to add resilience and useful redundancy to the financial system.

After the current crisis, we must ask questions about the resilience of our health- and social-care sectors and of the economy generally. How can we ensure that the economy can produce enough kit and people to meet the needs of health- and social-care systems? What other activities need to be ring-fenced to ensure that they are resilient to future extreme events, from viruses to cyberattacks and cyclones?

To answer these questions, we need highly granular data embedded in high-dimensional models of many interacting agents. Such agent-based models have been used extensively in the natural sciences, to study everything from ecosystems to galaxies. They have been used much less in understanding our economies.

Developing these models will not be easy. It calls for a concerted, bridge-building effort involving statisticians, physicists, epidemiologists, meteorologists, sociologists and economists. Intellectual cross-pollination of the kind that I (a policymaker) enjoyed with May (an ecologist) could, in time, help to contain the sorts of viral and economic contagion that are imposing such high costs on the world today.

BANK OF ENGLAND

© 2020 Springer Nature Limited. All rights reserved.

News in brief



PHOTOGRAPH: PHILIPPE CLEMENT/ARTERRA/UNIVERSAL IMAGES GROUP/GETTY; DATA SOURCE: *NATURE*

ARCHAEOLOGICAL WORKHORSE GETS MAJOR REBOOT

Radiocarbon dating — a key tool for determining the age of prehistoric samples — is due to be recalibrated using a slew of new data from around the world. The work combines thousands of data points from tree rings, lake and ocean sediments, corals and stalagmites, among other features, and means that the technique can be used to judge dates back to 55,000 years ago — 5,000 years further than the last calibration update, in 2013.

Archaeologists are downright giddy. "Maybe I've been in lockdown too long," tweeted Nicholas Sutton, an archaeologist at the University of Otago in Dunedin, New Zealand, "but … I'm really excited about it!"

The basis of radiocarbon dating is simple: all living things absorb carbon from the atmosphere and food sources around them, including a certain amount of natural, radioactive carbon-14. When the plant or animal dies, they stop absorbing, and the radioactive carbon that they've accumulated continues to decay. Measuring the amount left over gives an estimate as to how long something has been dead.

But this basic calculation assumes that the amount of carbon-14 in the environment has been constant in time and space — which it hasn't. In recent decades, the burning of fossil fuels and tests of nuclear bombs have radically altered the amount of carbon-14 in the air, and there are non-anthropogenic wobbles going much further back.

As a result, researchers have created conversion tables that match calendar dates with radiocarbon dates for different times and regions. Scientists are releasing new curves for the Northern Hemisphere (a model called IntCal20), Southern Hemisphere (SHCal20) and marine samples (MarineCal20). They will be published in the journal *Radiocarbon* in the next few months.

Tom Higham, an archaeological chronologist and director of the Oxford Radiocarbon Accelerator Unit, UK, says that recalibration is fundamental for understanding the chronology of hominins living 40,000 years ago, among other things. "I am really excited about calibrating our latest data using this curve."

VIRTUAL CONFERENCES FIND FAVOUR WITH SCIENTISTS

More than 80% of respondents to a *Nature* reader poll said that they would be in favour of some scientific conferences remaining virtual even after the coronavirus pandemic ends. Many meetings have been pushed online since March as a result of the global COVID-19 outbreak — including large, flagship conferences that usually attract thousands of attendees.

More than 40% of the roughly 500 survey respondents said that they had attended an online meeting (see 'Virtual reality').

Readers lauded some aspects of virtual meetings — in particular, improved accessibility, low costs and avoiding the hassle of travel. "I know colleagues around the world with limited budgets who've also suddenly been able to attend many more meetings," says Tom Brown, who studies energy-system modelling at Karlsruhe Institute of Technology in Germany. Some respondents found presentations to be clearer, liked that they could rewatch recorded talks, and felt that it is easier to speak up as an audience member using digital tools than it is in a meeting hall.

Downsides included clunky technology, connection issues and, most notably, the lack of serendipitous encounters, human interaction and socializing.

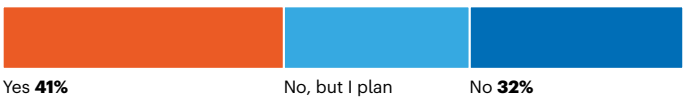
But many respondents thought that virtual conferences would improve in time and that the digital experience could help to make in-person meetings better, for instance through improving technologies that allow attendees to exchange data, knowledge and opinions during presentations.

Some advocated a hybrid model — a face-to-face meeting with increased virtual elements. "Some of the best parts of a conference, such as informal meetings, local flavour and more time to talk to speakers after sessions were completely lost. A hybrid system might recover some of these benefits," says Paul DeStefano, a physicist at Portland State University in Oregon.

VIRTUAL REALITY

In a *Nature* online poll, more than 40% of respondents said they had attended a scientific meeting held online because of the coronavirus pandemic. About 80% said some conferences should remain virtual even after the pandemic is over.

Have you attended a conference that was run virtually as a result of the pandemic? (499 respondents)



Do you think some meetings should continue to be virtual after the pandemic? (486 respondents)



Nature online poll run 20 April – 4 May 2020.

© 2020 Springer Nature Limited. All rights reserved.

News in focus



JUAN IGNACIO RONCORONI/EPA-EFE/SHUTTERSTOCK

Vaccines against the coronavirus are being tested in humans and animals.

CORONAVIRUS VACCINE TRIALS HAVE DELIVERED THEIR FIRST RESULTS — BUT THEIR PROMISE IS STILL UNCLEAR

Scientists urge caution over hints of success emerging from early human and animal studies.

By Ewen Callaway

As coronavirus vaccines hurtle through development, scientists are getting their first look at data that hint at how well different vaccines are likely to work. The picture, so far, is murky.

On 18 May, US biotech firm Moderna revealed the first data from a human trial: its COVID-19 vaccine triggered an immune response in people, and protected mice from lung infections with the coronavirus SARS-CoV-2. The results — which the company, based in Cambridge, Massachusetts, announced in a press release — were widely interpreted as positive and sent share prices surging. But some scientists say that, because the data haven't been published, they lack the details needed to properly evaluate those claims.

Tests of other fast-tracked vaccines show that they have prevented infections in the lungs of monkeys exposed to SARS-CoV-2 — but not in some other parts of the body. One — a vaccine being developed at the University of Oxford, UK, that is also in human trials — protected six monkeys from pneumonia, but the animals' noses harboured as much virus as did those of unvaccinated monkeys, researchers reported[1] last week in a bioRxiv preprint. A Chinese group reported similar caveats about its own vaccine's early animal tests this month[2].

Despite the uncertainties, all three teams are pressing ahead with clinical trials. These early studies in people are meant mainly to test safety, but larger clinical trials designed to determine whether the vaccines can actually protect humans from COVID-19 could report

© 2020 Springer Nature Limited. All rights reserved.

in the next few months.

Still, the early data together offer clues as to how coronavirus vaccines might generate a strong immune response. Scientists say that animal data will be crucial for understanding how coronavirus vaccines work, so that the most promising candidates can be identified quickly and then refined.

"We might have vaccines in the clinic that are useful in people within 12 or 18 months," says Dave O'Connor, a virologist at the University of Wisconsin–Madison. "But we're going to need to improve on them to develop second- and third-generation vaccines."

Immune response

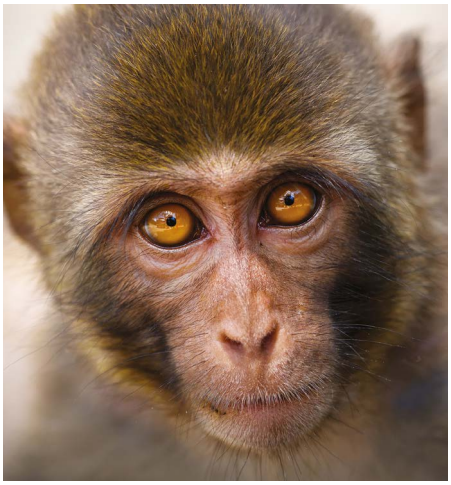
Moderna's vaccine, which is being co-developed with the US National Institute of Allergy and Infectious Diseases (NIAID) in Bethesda, Maryland, began safety testing in humans in March. The vaccine consists of messenger RNA instructions for building the coronavirus's spike protein; it causes human cells to churn out the foreign protein, alerting the immune system. Although such RNA-based vaccines are easy to develop, none has ever been licensed.

In its press release, the company reported that 45 study participants who received one or two doses of the vaccine developed a strong immune response to the virus. Researchers measured virus-recognizing antibodies in 25 participants, and detected levels similar to or higher than those found in the blood of people who have recovered from COVID-19.

Tal Zaks, Moderna's chief medical officer, said in a presentation to investors that these antibody levels bode well for the vaccine preventing infection. "If you get to the level of people who had disease, that should be enough."

But it's not at all clear whether the responses are enough to protect people from infection, because Moderna hasn't shared its data, says Peter Hotez, a vaccine scientist at Baylor College of Medicine in Houston, Texas. "I'm not convinced that this is really a positive result," Hotez says. He points to a 15 May bioRxiv preprint[3] that found that most people who have recovered from COVID-19 without hospitalization do not produce high levels of 'neutralizing antibodies', which block the virus from infecting cells. Moderna measured these potent antibodies in eight trial participants and found their levels to be similar to those in recovered patients.

Hotez also has doubts about the Oxford team's first results, which found that monkeys produced modest levels of neutralizing antibodies after receiving one dose of the vaccine (the same regime that is being tested in human trials). "It looks like those numbers need to be considerably higher to afford protection," says Hotez. The vaccine is made from a chimpanzee virus that has been genetically altered to produce a coronavirus protein.



MAGNUS LUNDGREN/WILD WONDERS OF CHINA/NATURE PICTURE LIBRARY

Rhesus macaques are used to test vaccines.

Hotez says that the vaccine being developed by Sinovac Biotech in Beijing seems to have elicited a more promising antibody response in macaque monkeys that received three doses, as reported[2] in a 5 May paper in *Science*. That vaccine is made of chemically inactivated SARS-CoV-2 particles.

No one yet knows the precise nature of the immune response that protects people from COVID-19, and the levels of neutralizing antibodies made by the monkeys in the Oxford study might be enough to protect people from infection, says Michael Diamond, a viral immunologist at Washington University in St. Louis, Missouri, who is a member of Moderna's scientific advisory board. If not, a second injection would probably boost levels. "What we don't know is how long they'll last," he adds.

"You want a vaccine that would protect against disease and against transmission."

Still more questions hover over experiments showing that vaccines can protect animals from infection. Moderna said its vaccine stopped the virus replicating in the lungs of mice. The rodents had been infected with a version of the virus that was genetically modified to let it attack mouse cells, which are not ordinarily susceptible to SARS-CoV-2, according to Zaks's presentation. But the mutation affects the protein that most vaccines, including Moderna's, use to stimulate the immune system; this could change the animals' response to infection[4].

The Oxford monkeys were given a high dose of virus after receiving the vaccine, says Sarah Gilbert, an Oxford vaccinologist who co-led the study with Vincent Munster, a virologist at NIAID's laboratories in Hamilton, Montana. This could explain why the vaccinated animals had just as much SARS-CoV-2 genetic material in their noses as did control animals, even though the vaccinated monkeys didn't develop any signs of pneumonia. Administering high doses ensures that the animals are infected with the virus, but it might not replicate natural infections. The Oxford study did not measure whether the virus was still infectious, Diamond says, and the genetic material could represent virus particles inactivated by the monkeys' immune response, or the viruses the researchers administered, rather than an ongoing infection.

Still, the result raises the possibility that vaccinated people could still spread the virus, says Douglas Reed, an aerobiologist at the University of Pittsburgh Center for Vaccine Research in Pennsylvania. "Ideally, you want a vaccine that would protect against disease and against transmission," he says.

Safety signs

Although assessing a vaccine's potential efficacy is difficult, the latest data are clearer on safety, say researchers. The Moderna vaccine caused few severe and no lasting health problems in trial participants. The vaccinated Oxford and Sinovac monkeys did not develop an exacerbated disease after infection – a key fear, because an inactivated vaccine for the related coronavirus that causes SARS (severe acute respiratory syndrome) showed signs of this in macaques[5].

None of the data should dissuade developers from doing human trials to determine whether the vaccines work, says Stanley Perlman, a coronavirologist at the University of Iowa in Iowa City.

Moderna will soon begin a phase II trial involving 600 participants. It hopes to begin a larger, phase III, efficacy trial in July, to test whether the vaccine can prevent disease in high-risk groups, such as health-care workers and people with underlying medical problems. Zaks said that further animal studies, including some in monkeys, were under way, and that it wasn't yet clear which animal would best predict whether and how the vaccine works.

The Oxford team has already enrolled more than 1,000 people in its UK trial. Some volunteers have received a placebo, so trials could determine whether the vaccine works in humans over the coming months. The lack of safety problems in the team's monkey study was reassuring, Gilbert says.

"We don't really need any more data from animal trials to continue," she says. "If we get human efficacy, we've got human efficacy, and that's what matters."

1. van Doremalen, N. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.05.13.093195 (2020).
2. Gao, Q. *et al. Science* https://doi.org/10.1126/science.abc1932 (2020).
3. Robbiani, D. F. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.05.13.092619 (2020).
4. Dinnon, K. H. III *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.05.06.081497 (2020).
5. Tseng, C.-T. *et al. PLoS ONE* **7**, e35421 (2012).

© 2020 Springer Nature Limited. All rights reserved.

ARE WOMEN PUBLISHING LESS DURING THE PANDEMIC? HERE'S WHAT THE DATA SAY

Early analyses suggest female academics are posting fewer preprints than men, and starting fewer projects.

By Giuliana Viglione

Quarantined with a six-year-old child underfoot, Megan Frederickson wondered how academics were managing to write papers during the COVID-19 pandemic. Lockdowns implemented to stem coronavirus spread meant that, overnight, many households worldwide had become an intersection of work, school and home life. Conversations on Twitter seemed to confirm Frederickson's suspicions: female academics, taking up increased childcare responsibilities, were falling behind their male peers at work.

But Frederickson, an ecologist at the University of Toronto, Canada, wanted to see what the data said. So, she looked at preprint servers to investigate whether women were posting fewer studies than they were before lockdowns began. The analysis — and several others — suggests that, across disciplines, women's publishing rate has fallen relative to men's amid the pandemic (see go.nature.com/2a5uwv5).

The results are consistent with the literature on the division of childcare between men and women, says Molly King, a sociologist at Santa Clara University in California. Evidence suggests that male academics are more likely to have a partner who does not work outside the home; their female colleagues, especially those in the natural sciences, are more likely to have a partner who is also an academic. Even in those dual-academic households, the evidence shows that women perform more household tasks than men do, she says. King suspects the same holds true for childcare.

Preprint analysis

In her analysis, Frederickson focused on the two preprint servers that she uses: the physical-sciences repository arXiv, and bioRxiv for the life sciences. To determine the genders of more than 73,000 authors named on 36,529 preprints, she compared the names with those in the US Social Security Administration's baby-name database, which registers the names and genders of children born in the United States.

Frederickson looked at arXiv studies posted between 15 March and 15 April in 2019 and in 2020. The number of women who authored preprints grew by 2.7% from 2019 to 2020 — but the number of male authors increased by 6.4% over that period. The increase in male authorship of bioRxiv preprints also outstripped that of female authorship, although by a smaller margin (see 'Preprint drop-off'). (The two servers are not directly comparable in Frederickson's analysis, because the program that she used pulled the names of only corresponding authors from bioRxiv, whereas all arXiv authors were included.)

"The differences are modest, but they're there," Frederickson says. She notes that lockdowns so far have been short compared with the usual research timeline, so the long-term effects on women's careers are still unclear.

The limitations of these types of name-based analysis are well known. Using names to predict gender can exclude non-binary people, and can misgender others. They are more likely to exclude authors with non-Western names. And, between disciplines, their utility can vary because of naming conventions — such as the use of initials instead of given names, as is common in astrophysics. Still, says Frederickson, over a large sample size, they can provide valuable insights into gender disparities in academia.

Fresh projects

Other researchers are finding similar trends. Cassidy Sugimoto, an information scientist at Indiana University Bloomington who studies gender disparities in research, conducted a separate analysis of author gender on nine popular preprint servers (see go.nature.com/2xhxqxr). Methodological differences meant that the two analyses are not directly comparable, but Frederickson's work "converges with what we're seeing", says Sugimoto.

Sugimoto points out that the preprints being published even now probably rely on labour that was performed many months ago. "The scientific publication process doesn't lend itself to timely analyses," she says. So her study also included databases that log registered reports, which indicate the initiation of new research projects.

In 2 of the 3 registered-report repositories, covering more than 14,000 reports with authors whose genders could be matched, Sugimoto's team found a decrease in the proportion of submissions by female principal investigators from March and April of 2019 to the same months in 2020, when lockdowns started. They also saw a declining proportion of women publishing on several preprint servers, including EarthArXiv and medRxiv. These differences were more pronounced when looking at first authors, who are usually early-career researchers, than at last authors, who are often the most senior faculty members on a study.

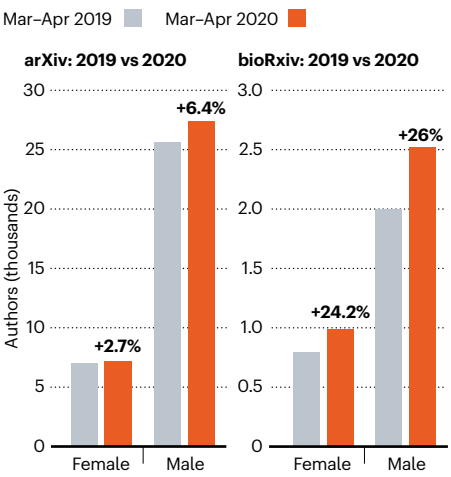
"This is what's the most worrying to me, because those consequences are long term," Sugimoto says. "The best predictor of a publication is a previous publication."

PREPRINT DROP-OFF

Two separate analyses show that women's posting rate on preprint servers has slowed during the coronavirus pandemic.

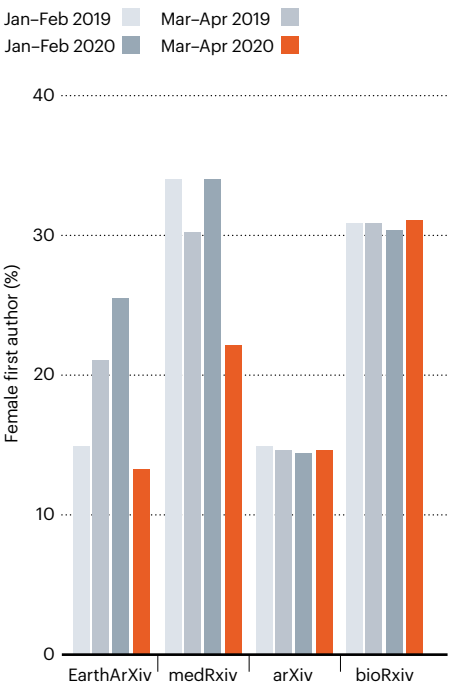
All-author analysis

When compared with March and April 2019, the number of male authors on preprints posted to bioRxiv and arXiv has grown faster than the number of female authors in that period this year.



First-author analysis

At many preprint servers, women were submitting at a lower rate in March and April, as compared with the preceding two months and the same months of the previous year.

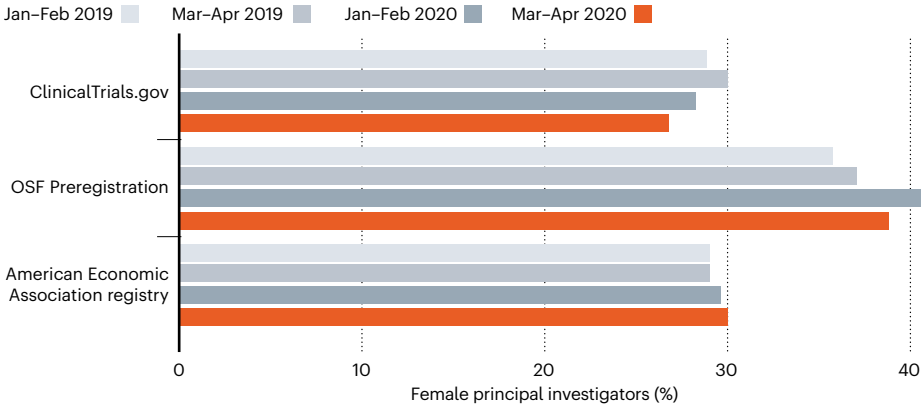


SOURCES: ALL-AUTHOR ANALYSIS, M. FREDERICKSON HTTPS://GO.NATURE.COM/2A5UWV5 (2020); FIRST-AUTHOR ANALYSIS, P. VINCENT-LAMARRE, C. R. SUGIMOTO AND V. LARIVIÈRE *NATURE INDEX* HTTPS://GO.NATURE.COM/2XHXQXR (2020)

© 2020 Springer Nature Limited. All rights reserved.

FEWER NEW PROJECTS

Women are registering a smaller proportion of research projects than before the pandemic, according to an analysis of registered-report repositories.



SOURCE: P. VINCENT-LAMARRE, C. R. SUGIMOTO AND V. LARIVIÈRE *NATURE INDEX* HTTPS://GO.NATURE.COM/2XHXQXR (2020)

In economics, too, there are indications that the pandemic is disproportionately affecting younger researchers, says Noriko Amano-Patiño, an economist at the University of Cambridge, UK. Taken as a whole, there aren't clear discrepancies in the overall number of working papers — a preprint-like publication format in economics — that have been submitted to three major repositories, or in invited commentaries submitted to a fourth site that publishes research-based policy analyses (see 'Fewer new projects').

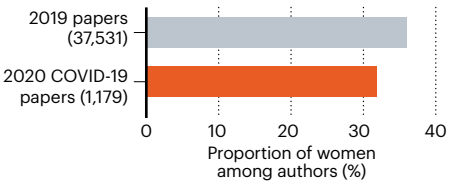
Academic responsibilities

Amano-Patiño and her collaborators also examined who was working on pandemic-related research questions using a COVID-19-specific repository (see go.nature.com/36sj2cm). Although women have consistently authored about 20% of working papers since 2015, they make up only 12% of the authors of new COVID-19-related research. Amano-Patiño suspects that, in addition to their childcare responsibilities, early- and mid-career researchers, especially women, might be more risk-averse and thus less likely to jump into a new field of research. "Mostly senior economists are taking their bite into these new areas," says Amano-Patiño. "And junior women are the ones that seem to be missing out the most."

"Unfortunately, these findings are not surprising," says Olga Shurchkov, an economist at Wellesley College in Massachusetts. Shurchkov came to similar conclusions in a separate analysis of economists' productivity during the pandemic (see go.nature.com/2zyuebi). And a 13 May arXiv preprint (J. P. Andersen *et al.* Preprint at https://arxiv.org/abs/2005.06303; 2020) shows the same trends in pandemic-related medical literature (see 'COVID-19 effect'). Compared with the proportion of women among authors of nearly 40,000 articles published in US medical journals in 2019, the proportion on COVID-19 papers has dropped by 16%.

COVID-19 EFFECT

An analysis that looked at 13 medical journals found that the proportion of female authors for COVID-19 papers is lower than the average for all studies published in 2019 in the journals.



SOURCE: J. P. ANDERSEN *ET AL.* PREPRINT AT HTTPS://ARXIV.ORG/ABS/2005.06303 (2020)

Increased childcare responsibility is one issue. In addition, women are more likely than men to take care of ailing relatives, says Rosario Rogel-Salazar, a sociologist at the Autonomous University of Mexico State in Toluca. These effects are probably exacerbated in the global south, she notes, because women there have more children on average than do their northern counterparts.

And women face other barriers to productivity. Female faculty members, on average, shoulder more teaching responsibilities, so the sudden shift to online teaching — and the curriculum adjustments that it requires — disproportionately affects women, King says. And because many institutions are shut, non-research university commitments — such as participation in hiring and curriculum committees — are probably taking up less time. These are often dominated by senior faculty members, more of whom are men. As a result, men could find themselves with more time to write papers.

Because these effects will be compounded as lockdowns persist, universities and funders should take steps to mitigate gender disparities as quickly as possible, Shurchkov says. "They point to a problem that, if left unaddressed, can potentially have grave consequences for diversity in academia."

THE QUEST TO ADDRESS INEQUALITY DURING THE PANDEMIC

Better data, testing and preparedness could reduce COVID-19's outsized toll on people of colour.

By Nidhi Subbaraman

As figures emerge on the disproportionate toll that COVID-19 is taking on people of colour in the United States, scientists are suggesting measures to help mitigate the inequalities.

They say that better data are needed on the incidence of the disease, that testing needs to be ramped up and that hospitals serving people at risk need to prepare more effectively. Researchers and some US lawmakers are now calling for a national commission devoted to identifying racial disparities in health that would act as a unified voice in trying to overcome them.

The unequal impact of the coronavirus is not solely a US problem, but the disparities are harshly felt in the United States, which currently has the highest number of COVID-19 infections and deaths in the world.

The US Centers for Disease Control and Prevention (CDC) started releasing death and infection rates broken down by race and ethnicity in late April, after a public outcry from lawmakers, doctors and civil-rights groups.

The breakdowns were available for just 35% of US deaths. But as these and other data start to come in, they paint a stark picture of disproportionate disease burden.

Many of the causes of these health disparities are systemic and well known. "We're getting infected more because we are exposed more and less protected," says Camara Phyllis Jones, an epidemiologist at the Rollins School of Public Health at Emory University in Atlanta, Georgia. Existing socio-economic and health disparities can at least partially explain why people of colour are getting ill and dying at disproportionate rates.

In many parts of the United States, people of

© 2020 Springer Nature Limited. All rights reserved.



ETIENNE LAURENT/EPA-EFE/SHUTTERSTOCK

A woman has her temperature taken in Compton, a city south of Los Angeles, California.

colour make up a high proportion of those in some low-paid professions that have elevated risks of exposure to the virus — people who staff grocery stores, drive buses and work at food plants, for example. Also, COVID-19 is deadlier for people with chronic conditions, including diabetes, obesity and cardiovascular disease. These have a higher incidence in many minority ethnic and racial groups in the United States.

Reaching for solutions

At the earliest stages of the epidemic, aspects of the US response might have made things worse, says Enrique Neblett, a psychologist who studies race and health at the University of Michigan in Ann Arbor. With limited tests available, US authorities initially reserved them for people with symptoms who had a recent history of overseas travel. This could have excluded people from disadvantaged socio-economic backgrounds, including people of colour, says Neblett. "By having that as a criterion, they're automatically less likely to be tested." Looking ahead, adapting the approaches to getting tests to the people and communities that are most at risk should be a priority, he says.

For example, in Louisiana — one of the first states to report data by race and ethnic group — testing teams went to poorer neighbourhoods to reach people without cars, who would have trouble making it to drive-through testing sites.

Outreach must extend beyond testing, to all aspects of the response, says Evelynn Hammonds, a historian of medicine at Harvard University in Cambridge, Massachusetts. Clinical trials, for example, must actively work to recruit a diverse population, otherwise the treatments and vaccines might not be equally effective. "We already know there's a real problem with making sure that populations that are enlisted into clinical trials need to be diverse," Hammonds says.

And hospitals in neighbourhoods that are likely to see a surge in cases because they serve more severely affected groups should be equipped sufficiently, says Jones, who was president of the American Public Health Association in 2016. "If we know that these neighbourhoods are the ones being adversely impacted, then we need to move the ventilators and the health staff there," Jones says.

A coherent voice

Billions of federal dollars have been deployed in the United States to tackle the pandemic, but because such emergencies require a coordinated response across national agencies, some have argued for a central commission that will represent the needs of minority racial and ethnic groups.

"The call is for a national coherent voice that starts to talk about these issues," says Cato Laurencin, an orthopaedic surgeon and biomedical engineer at the University of Connecticut in Farmington, who led a roundtable discussion on diversity at the US National Academies of Science, Engineering, and Medicine.

Hammonds says that such a group could be effective because local US leaders typically prescribe public-health guidance and decisions, and so far have differed in their response to the pandemic. She compared New York's rapid yet measured response with that in Georgia, where the governor announced a gradual re-opening of the economy in mid-April — despite advice against the move from public-health experts. This pandemic could be an opportunity to deploy consistent attention towards the needs of under-served communities, she says. "If there is a positive to come out of this, that would be one of them."

Q&A

Astronaut with sights on Mars



BILL STAFFORD/NASA

Jessica Watkins graduated as a member of NASA's newest astronaut class in January. As a planetary geologist, she is a leading candidate to participate in the agency's Artemis programme, which aims to send people back to the Moon by the end of 2024. Further down the line, there might even be a trip to Mars, which she studied during her PhD. *Nature* spoke to Watkins about her career.

Why did you join the astronaut corps?
I have wanted to be an astronaut since I was little. There was something that always pulled me towards space — the idea of exploration, of wanting to push boundaries and capabilities, both technically and physically, but also mentally and spiritually. I kind of stumbled into geology and fell in love with that. And then the stars aligned for me to end up here.

What's your favourite planet?
Mars is definitely my first love. I remember writing a book about a Martian in fifth grade. What intrigued me the most about Mars is how Earth-like it is, and how we're able to use Earth as an analogue to understand more about Mars. Now, given the direction that NASA is going in — we're talking about going back to the Moon in 2024, through Artemis — the Moon has become a significant interest for me, as well. I'm definitely brushing up on lunar geology and what it's going to be like on the surface.

How can space exploration inspire us during the current public-health crisis?
This pandemic is asking us to band together as humans, to do the right thing to help save each other. There's something really analogous to human spaceflight in that. Human spaceflight is about humans pursuing hard things, doing it together, and doing it in spite of differences that we may have created. Having that perspective allows you to see Earth for what it is. It's one body. We're all in this together.

Interview by Alexandra Witze
This interview has been edited for length and clarity.

© 2020 Springer Nature Limited. All rights reserved.



ADAM WALTZ/ABC15 ARIZONA

Protesters rallying in Arizona against lockdowns held up signs carrying anti-vaccine messages and promoting unproven treatments.

BATTLING THE INFODEMIC

Researchers are analysing false rumours and disinformation about COVID-19 in hopes of curbing their spread. **By Philip Ball and Amy Maxmen**

In the first few months of 2020, wild conspiracy theories about Bill Gates and the new coronavirus began sprouting online. Gates, the Microsoft co-founder and billionaire philanthropist who has funded efforts to control the virus with treatments, vaccines and technology, had himself created the virus, argued one theory. He had patented it, said another. He'd use vaccines to control people, declared a third. The false claims quietly proliferated among groups predisposed to spread the message — people opposed to vaccines, globalization or the privacy infringements enabled by technology. Then one went mainstream.

On 19 March, the website Biohackinfo.com falsely claimed that Gates planned to use a coronavirus vaccine as a ploy to monitor people through an injected microchip or quantum-dot spy software. Two days later, traffic started flowing to a YouTube video on the idea. It's been viewed nearly two million times. The idea reached Roger Stone — a former adviser to US President Donald Trump — who in April discussed the theory on a radio show, adding that he'd never trust a coronavirus vaccine that Gates had funded. The interview was covered by the newspaper the *New York Post*, which didn't debunk the notion. Then that article was liked, shared or commented on by nearly one million people on Facebook. "That's better performance than most mainstream media news stories," says Joan Donovan, a sociologist at Harvard University in Cambridge, Massachusetts.

Donovan charts the path of this piece of disinformation like an epidemiologist tracking the transmission of a new virus. As with epidemics, there are 'superspreader' moments. After the *New York Post* story went live, several high-profile figures with nearly one million Facebook followers each posted their own alarming comments, as if the story about Gates devising vaccines to track people were true.

The Gates conspiracy theories are part of an ocean of misinformation on COVID-19 that is spreading online. Every major news event comes drenched in rumours and propaganda.

© 2020 Springer Nature Limited. All rights reserved.

But COVID-19 is "the perfect storm for the diffusion of false rumour and fake news", says data scientist Walter Quattrociocchi at the Ca'Foscari University of Venice, Italy. People are spending more time at home, and searching online for answers to an uncertain and rapidly changing situation. "The topic is polarizing, scary, captivating. And it's really easy for everyone to get information that is consistent with their system of belief," Quattrociocchi says. The World Health Organization (WHO) has called the situation an infodemic: "An over-abundance of information — some accurate and some not — rendering it difficult to find trustworthy sources of information and reliable guidance."

For researchers who track how information spreads, COVID-19 is an experimental subject like no other. "This is an opportunity to see how the whole world pays attention to a topic," says Renée diResta at the Stanford Internet Observatory in California. She and many others have been scrambling to track and analyse the disparate falsehoods floating around — both 'misinformation', which is wrong but not deliberately misleading, and 'disinformation', which refers to organized falsehoods that are intended to deceive. In a global health crisis, inaccurate information doesn't only mislead, but could be a matter of life and death if people start taking unproven drugs, ignoring public-health advice, or refusing a coronavirus vaccine if one becomes available.

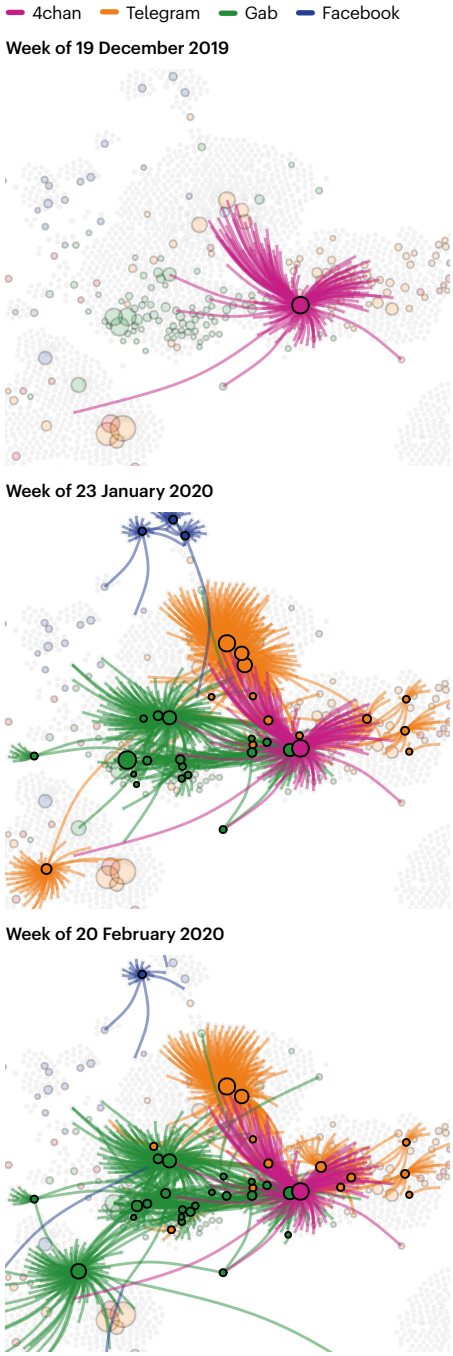
By studying the sources and spread of false information about COVID-19, researchers hope to understand where such information comes from, how it grows and — they hope — how to elevate facts over falsehood. It's a battle that can't be won completely, researchers agree — it's not possible to stop people from spreading ill-founded rumours. But in the language of epidemiology, the hope is to come up with effective strategies to 'flatten the curve' of the infodemic, so that bad information can't spread as far and as fast.

No filter

Researchers have been monitoring the flow of information online for years, and have a good sense of how unreliable rumours start and spread. Over the past 15 years, technology and shifting societal norms have removed many of the filters that were once placed on information, says Amil Khan, director of the communications agency Valent Projects in London, who has worked on analysing misinformation for the UK government. Rumour-mongers who might once have been isolated in their local communities can connect with like-minded sceptics anywhere in the world. The social-media platforms they use are run to maximize user engagement, rather than to favour evidence-based information. As these platforms have exploded in popularity over the past decade and a half, so political partisanship and voices that distrust authority have grown too.

HIGHWAYS OF HATE

Neil Johnson at George Washington University in Washington DC and his team mapped how malicious content about a pneumonia-like disease, possibly COVID-19, started on the forum 4chan in December. By January, the content had spread to other social-media platforms — Gab, Telegram and Facebook — through links connecting pages on one platform with another.



SOURCE: REF. 4

To chart the current infodemic, data scientists and communications researchers are now analysing millions of messages on social media. A team led by Emilio Ferrara, a data scientist at the University of Southern California in Los Angeles, has released a data set of more than 120 million tweets on the coronavirus[1]. Theoretical physicist Manlio De Domenico at the Bruno Kessler Institute, a research institute for artificial intelligence in Trento, Italy, has set up what he calls a COVID-19 "infodemic observatory", using automated software to watch 4.7 million tweets on COVID-19 streaming past every day. (The actual figure is higher, but that is as many as Twitter will allow the team to track.) De Domenico and his team evaluate the tweets' emotional content and, where possible, the region they were sent from. They then estimate their reliability by looking at the sources to which a message links. (Like many data scientists, they rely on the work of fact-checking journalists to distinguish reliable news sources or claims from unreliable ones.) Similarly, in March, Quattrociocchi and his co-workers reported[2] a data set of around 1.3 million posts and 7.5 million comments on COVID-19 from several social-media platforms, including Reddit, WhatsApp, Instagram and Gab (known for its right-wing audience), from 1 January to mid-February.

A study in 2018 suggested that false news generally travels faster than reliable news on Twitter[3]. But that isn't necessarily the case in this pandemic, says Quattrociocchi. His team followed some examples of false and true COVID-19 news — as classified by fact-checker sites — and found that reliable posts saw as many reactions as unreliable posts on Twitter[2]. The analysis is preliminary and hasn't yet been peer reviewed.

Ferrara says that in the millions of tweets about the coronavirus in January, misinformation didn't dominate the discussion. Much of the confusion at the start of the pandemic related to fundamental scientific uncertainties about the outbreak. Key features of the virus — its transmissibility, for instance, and its case-fatality rate — could be estimated only with large error margins. Where expert scientists were honest about this, says biologist Carl Bergstrom at the University of Washington in Seattle, it created an "uncertainty vacuum" that allowed superficially reputable sources to jump in without real expertise. These included academics with meagre credentials for pronouncing on epidemiology, he says, or analysts who were good at crunching numbers but lacked a deep understanding of the underlying science.

Politics and scams

As the pandemic shifted to the United States and Europe, false information increased, says Donovan. A sizeable part of the problem has been political. A briefing prepared for the European Parliament in April alleged that Russia and China are "driving parallel information campaigns, conveying the overall message that democratic state actors are failing and that European citizens cannot trust their health systems, whereas their authoritarian systems can save the world."

© 2020 Springer Nature Limited. All rights reserved.

The messages of US President Donald Trump and his administration are sowing their own political chaos. This includes Trump's insistence on referring to the 'Chinese' or 'Wuhan' coronavirus and his advocacy of unproven (and even hazardous) 'cures', and the allegation by US Secretary of State Mike Pompeo that the virus originated in a laboratory, despite the lack of evidence.

There are organized scams, too. More than 68,000 website domains have been registered this year with keywords associated with the coronavirus, says Donovan. She's reviewed ones that sell fake treatments for COVID-19, and others that collect personal information. Google's search-engine algorithms rank information from the WHO and other public-health agencies higher than that from other sources, but rankings vary depending on what terms a person enters in a search. Some scam sites have managed to come out ahead by using a combination of keywords optimized and targeted to a particular audience, such as newly unemployed people, Donovan says.

Spreading agendas

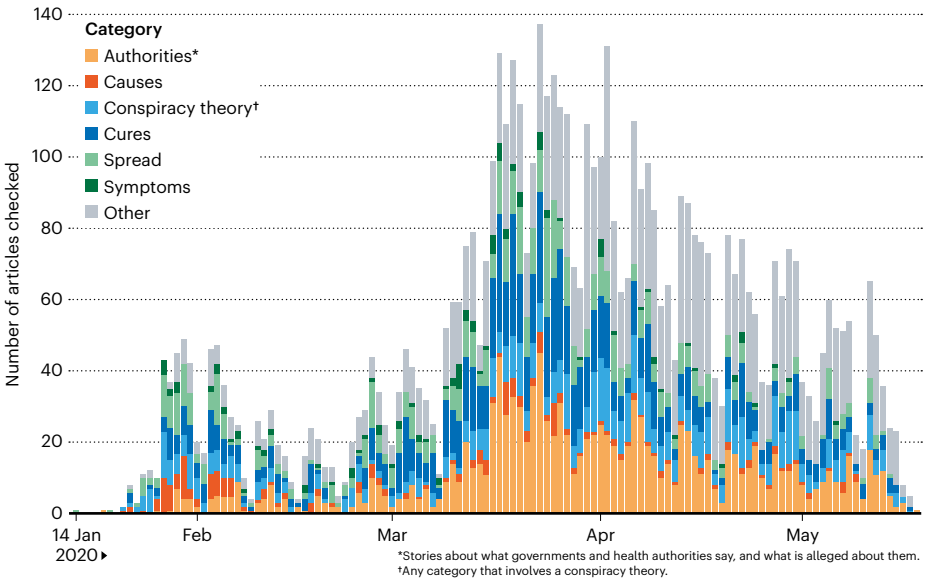
Many of the falsehoods online don't have obvious sources or intentions. Rather, they often begin with niche groups mobilizing around their favoured agendas. Neil Johnson, a physicist at George Washington University in Washington DC, has reported[4] COVID-19 misinformation narratives taking shape among online communities of extremist and far-right 'hate' groups, which occupy largely unregulated platforms including VKontakte, Gab and 4Chan, as well as mainstream ones such as Facebook and Instagram.

The study says that a "hate multiverse" is exploiting the COVID-19 pandemic to spread racism and other malicious agendas, focusing an initially rather diverse and incoherent set of messages into a few dominant narratives, such as blaming Jews and immigrants for starting or spreading the virus, or asserting that it is a weapon being used by the "Deep State" to control population growth (see 'Highways of hate').

An alarming feature of this network is its capacity to draw in outside users through what Johnson and his team call "wormhole" links. These are shortcuts from a network engaged with quite different issues. The hate multiverse, the researchers say, "acts like a global funnel that can suck individuals from a mainstream cluster on a platform that invests significant resources in moderation, into less moderated platforms like 4Chan or Telegram". As a result, Johnson says, racist views are starting to appear in the anti-vaccine communities, too. "The rise of fear and misinformation around COVID-19 has allowed promoters of malicious matter and hate to engage with mainstream audiences around a common topic of interest, and potentially push them toward hateful views," his team says in the paper.

A FACT-CHECKING FRENZY

Fact-checkers have worked overtime correcting COVID-19 falsehoods. One alliance has collated more than 6,000 examples of fact-checks across a broad range of categories since 14 January. Data as of 19 May.



*Stories about what governments and health authorities say, and what is alleged about them.
†Any category that involves a conspiracy theory.

SOURCE: IFCN/CORONAVIRUSFACTSALLIANCE

Dangerous spread

As misinformation grows, it sometimes becomes deadly. On Twitter in early March, technology entrepreneurs and investors shared a document prematurely extolling the benefits of chloroquine, an old malaria drug, as an antiviral against COVID-19. The document, which claimed that the drug had produced favourable outcomes in China and South Korea, was widely passed around even before the results of a small, non-randomized French trial of the related drug hydroxychloroquine[5] were posted online on 17 March. The next day, Fox News aired a segment with one of the authors of the original document. And the following day, Trump called the drugs "very powerful" at a press briefing, despite the lack of evidence. There were small spikes in Google searches for hydroxychloroquine, chloroquine and their key ingredient, quinine, in mid-March — with the largest surge on the day of Trump's remarks, Donovan found using Google Trends. "Just like toilet paper, masks and hand sanitizer, if there was a product to be had, it would have sold out," she says. Indeed, it did in some places, worrying people who need the drugs to treat conditions such as lupus. Hospitals have reported poisonings in people who experienced toxic side effects from pills containing chloroquine, and such a large number of people with COVID-19 have been asking for the drug that it has derailed clinical trials of other treatments.

"We are producing much more information than what we can really parse and consume."

Fox News has been particularly scrutinized for its part in amplifying dangerous misinformation. In a phone survey of 1,000 randomly chosen Americans in early March[6], communication researchers found that respondents who tended to get their information from mainstream broadcast and print media had more accurate ideas about the disease's lethality and how to protect themselves from infection than did those who got their news mostly from conservative media (such as Fox News and Rush Limbaugh's radio show) or from social media. That held true even after factors such as political affiliation, gender, age and education were controlled for.

Those results echo another study, as yet not peer reviewed, in which economists at the University of Chicago in Illinois tried to analyse the effects of two Fox News presenters on viewers' opinions during February, as the coronavirus began to spread beyond China. One presenter, Sean Hannity, downplayed the coronavirus's risk and accused Democrats of using it as a weapon to undermine the president; the other, Tucker Carlson, reported that the disease was serious. The study found that areas of the country where more viewers watched Hannity saw more cases and deaths than did those where more watched Carlson — a divergence that disappeared when Hannity adjusted his position to take the pandemic more seriously.

De Domenico says he is encouraged that, as the crisis has deepened, so has many people's determination to find more reliable information. "When COVID-19 started to hit each country, we have observed dramatic changes of attitude," he says. "People started to consume and share more reliable news from

© 2020 Springer Nature Limited. All rights reserved.

trusted sources." Of course, the goal is to have people listening to the best available advice on risk before they watch people die around them, Donovan says.

Flattening the curve

In March, Brazilian President Jair Bolsonaro began to spread misinformation on social media — posting a video that falsely said hydroxychloroquine was an effective treatment for COVID-19 — but was stopped in his tracks. Twitter, Facebook and YouTube took the unprecedented step of deleting posts from a head of state, on the grounds that they could cause harm.

Social-media platforms have stepped up their efforts to flag or remove misinformation and to guide people to reliable sources. In mid-March, Facebook, Google, LinkedIn, Microsoft, Reddit, Twitter and YouTube issued a joint statement saying that they were working together on "combating fraud and misinformation about the virus". Facebook and Google have banned advertisements for 'miracle cures' or overpriced face masks, for example. YouTube is promoting 'verified' information videos about the coronavirus.

Social-media platforms often rely on fact-checkers at independent media organizations to flag up misleading content. In January, 88 media organizations around the world joined together to record their fact-checks of COVID-19 claims in a database maintained by the International Fact-checking Network (IFCN), part of the Poynter Institute for Media Studies in St Petersburg, Florida (see 'A fact-checking frenzy'). The database currently holds more than 6,000 examples, and the IFCN is now inviting academics to dig into the data. (Another site, Google's fact-check explorer, records more than 2,700 fact-checks about COVID-19.) But some fact-checking organizations, such as Snopes, have admitted to being overwhelmed by the quantity of information they are having to deal with. "The problem with infodemics is its huge scale: collectively, we are producing much more information than what we can really parse and consume," says De Domenico. "Even having thousands of professional fact-checkers might not be enough."

Communication scholar Scott Brennen at the Oxford Internet Institute, UK, and his co-workers have found that social-media companies have done a decent job of removing misleading posts, given the hard task. The team followed up 225 pieces of misinformation about the coronavirus that independent fact-checkers had collated in the IFCN or Google databases as false or misleading. In a 7 April report, the team found that by the end of March, only some 25% of these false claims remained in place without warning labels on YouTube and Facebook, although on Twitter that proportion was 59% (see go.nature.com/2tvhuj5). And Ferrara says that about 5% of the 11 million Twitter users his team has studied so far in its COVID-19 database have been shut down for violating the platform's policies of use, and that these tended to be unusually active accounts.

But some creators of content have found ways to delay detection by social-media moderators, Donovan notes, in what she calls "hidden virality". One way is to post content in private groups on Facebook. Because the platform relies largely on its users to flag up bad information, shares of misleading posts in private communities are flagged less often because everyone in the group tends to agree with one another, she says. Donovan used to study white supremacy online, and says a lot of 'alt-right' content wasn't flagged until it leaked into public Facebook domains. Using CrowdTangle, a social-media-tracking tool owned by Facebook, Donovan found that more than 90% of the million or so interactions referring to the *New York Post* article about the Gates vaccine conspiracy were on private pages.

Another way in which manipulators slip past moderation is by sharing the same post from a new location online, says Donovan. For instance, when people on Facebook began sharing an article that alleged that 21 million people had died of COVID-19 in China, Facebook put a label on the article to indicate that it contained dubious information, and limited its ranking so that it wasn't prioritized in a search (China has confirmed many fewer deaths: 4,638). Immediately, however, people began posting a copy of the article that had been stored on the Internet Archive, a website that preserves content. This copy was shared 118,000 times before Facebook placed a warning on the link.

"The problem isn't a lack of facts. It's about what sources people trust."

Quattrociocchi says that, faced with regulation of content on platforms such as Twitter and Facebook, some misinformation simply migrates elsewhere: regulation is currently worse, he says, on Gab and WhatsApp. And there is only so much you can do to police social media: "If someone is really committed," says Ferrara, "once you suspend them, they go back and create another account."

Donovan agrees, but argues that social-media companies could implement stronger, faster moderation, such as finding when posts that have already been flagged, or deleted, are revived with alternative links. In addition, she says, social-media firms might need to adjust their policies on permitting political discourse when it threatens lives. She says that health misinformation is increasingly being buried in messages that seem strictly political at first glance. A Facebook group urging protests against stay-at-home restrictions — Re-Open Alabama — featured a video (viewed 868,000 times) of a doctor saying that his colleagues have determined that COVID-19 is similar to influenza, and "it shows healthy people don't need to shelter in place anymore". Those messages could lead people to ignore public-health guidance and endanger many others, says Donovan. But Facebook has been slow to curb these messages because they seem to be expressing political opinions. "It's important to demonstrate to platform companies that they aren't moderating political speech," Donovan says. "They need to look at what kind of health misinformation backs their claims that restrictions are unjustified." (Facebook did not reply to a request for comment.)

Gaining trust

Efforts to raise the profile of good information, and slap a warning label on the bad, can only go so far, says DiResta. "If people think the WHO is anti-American, or Anthony Fauci is corrupt, or that Bill Gates is evil, then elevating an alternative source doesn't do much — it just makes people think that platform is colluding with that source," she says. "The problem isn't a lack of facts, it's about what sources people trust."

Brennen agrees. "The people in conspiracy communities think that they are doing what they should: being critical consumers of media," he says. "They think they are doing their own research, and that what the consensus might advocate is itself misinformation."

That sentiment could grow if public-health authorities don't inspire confidence when they change their advice from week to week — on facemasks, for example, or on immunity to COVID-19. Some researchers say the authorities could be doing a better job at explaining the evidence, or lack of it, that guided them.

For now, US polling suggests that the public still supports vaccines. But anti-vaccine protesters are making more noise. At rallies protesting against lockdowns in California in May, for instance, some protestors carried signs saying, "No Mandatory Vaccines". Anti-vaccination online hubs are leaping on to COVID-19, says Johnson. "It's almost like they've been waiting for this. It crystallizes everything they've been saying."

Philip Ball is a science writer based in London, and **Amy Maxmen** is a senior reporter at *Nature*, based in San Francisco, California.

1. Chen, E., Lerman, K. & Ferrara, E. Preprint at https://arxiv.org/abs/2003.07372 (2020).
2. Cinelli, M. *et al*. Preprint at https://arxiv.org/abs/2003.05004 (2020).
3. Vosoughi, S., Roy, D. & Aral, S. *Science* **359**, 1146–1151 (2018).
4. Velásquez, N. *et al*. Preprint at https://arxiv.org/abs/2004.00673 (2020).
5. Gautret, P. *et al*. *Int. J. Antimicrob. Agents* https://doi.org/10.1016/j.ijantimicag.2020.105949 (2020).
6. Jamieson, K. H. & Albarracín, D. *Harvard Kennedy Sch. Misinform. Rev.* https://doi.org/10.37016/mr-2020-012 (2020).

© 2020 Springer Nature Limited. All rights reserved.

Books & arts



ZABED HASNAIN CHOWDHURY/SOPA IMAGES/LIGHTROCKET VIA GETTY

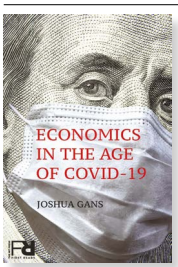
Garment workers in Bangladesh risk their lives in an industry devastated by the collapse in global spending.

COVID economics — first of many books

Breakneck triage nails many diagnoses. Deeper treatment is needed. **By Philip Ball**

There has never been a harder time to be a political leader. The choices that must be made are enormous, the consequences potentially catastrophic, the science guiding those decisions uncertain — and there is no precedent. As a result, the COVID-19 pandemic has revealed some of the best and the worst in the world's leaders: from opportunism and denial to compassion and clarity.

It's a shame that policymakers did not have books such as Joshua Gans's *Economics in the Age of COVID-19* to lay out the issues for them in January. It is remarkable that they do already. Gans completed this book at breakneck speed, by late March. His attempt to explain the economic thinking that should guide policy is useful, but inevitably limited.



Economics in the Age of COVID-19
Joshua Gans
MIT Press (2020).

With the situation and knowledge changing daily, unfurling events will always render some aspects of such an analysis obsolete. In this sense, Gans, an economist at the University of Toronto in Canada, has taken a brave shot at an impossible task. Ultimately, economic thinking will need wider horizons.

The crisis has forced some politicians, especially on the right, to go against deeply held inclinations by implementing interventions and financial handouts that, in normal times, even most of their opponents would deem excessive. Countries have tried to freeze their economies and prop up the absence

© 2020 Springer Nature Limited. All rights reserved.



USDA/ALAMY

There has been a surge in COVID-19 cases among meatpackers in the United States.

of liquidity and wages with eye-watering subsidies until the wheels start turning again.

Therein lies the difference from the oft-cited comparison with wartime economics. In that situation, activity continues, but redirected. The present worldwide lockdowns have drastically shrunk the workforce. Aside from essential workers — in health, care, food and transport, say — only those jobs that can be done alone from home can safely continue (never have I felt luckier to be a writer). This has sometimes been presented, too simplistically, as creating a choice between saving lives or saving the economy. As many countries have now passed the (first?) peak of infections, discussion has turned to the dangers to health posed by an economy left too long in stasis.

That discussion needs to happen, but it risks becoming facile, too. Presenting lives versus livelihoods as a dichotomy is used in defence of leaders who hesitated to impose a lockdown. That, Gans shows, is mistaken. The highly infectious nature and the fatality rate of COVID-19, which were both clear early on, even if exact numbers were not, meant there was never a gradual trade-off to be had: a dash more economy at the price of a few more deaths. "If you know you are going to shut down the country eventually, there are huge returns to doing it quickly," Gans writes. It is the only way to keep choices open as more is learnt about the virus and its spread.

This is not hindsight: Gans was writing while the UK and US governments were procrastinating. Nor is it just about saving lives in the short term. "Pursuing public health can be consistent with superior long-run economic performance," Gans writes.

And to be effective, that decision to shutter must be made with "resolve, clarity [and] transparency". If leaders downplay the enormity of the crisis, prevaricate or issue weak behavioural guidelines — rather than expectations with consequences — then individuals will "do as they often do and pursue their own interest", and will "keep businesses open and keep engaging in social life".

"How do you keep the economy in suspended animation, without the onset of necrosis?"

Then there's the question of how to manage the crisis in an economy on pause. Again, ideology might clash with reality. If you urgently need masks or ventilators, then there's no time to put it out to tender and let market mechanisms make the choices about who gets the contract and the product. There must be centralized decision-making and allocation, even if that risks a degree of 'inefficiency'.

And how do you keep the economy in suspended animation, without the onset of necrosis? Governments have generally realized that they must help to cover lost wages, but the details are very tricky. The options of providing financial assistance to cover bills, such as rent and mortgages, suspending those costs or covering them directly are not equivalent. Gans says that the aim must be for payments, subsidies and loans to "ensure that people's short-term disruptions are not translated into long-term breakups". One solution, he suggests, is repayment of government loans over time through taxation.

Is the past any guide? Gans touches on the only comparable event in recent times, the 1918 influenza pandemic. The economic consequences of that were complicated because it occurred directly after a global war (mobilization of troops exacerbated that outbreak). Gans might also have mentioned the AIDS epidemic in Africa, which has in some regions been devastating enough to orphan generations, deplete the workforce and hamper economic development. Prioritizing the economy over health is not necessarily taking the long view.

Gans does a good job fleshing out the requirements for an exit strategy. He says we

© 2020 Springer Nature Limited. All rights reserved.

now need to "invest in the testing economy", for example to establish who can safely return to work and to monitor workplace safety. "Countries and regions that were able to test, trace, and then isolate the infected were able to contain the virus quickly and reopen their economies sooner", he points out — in a ringing endorsement of World Health Organization policy. Even with a vaccine, he says, testing is likely to be a part of our daily lives for many years.

He also offers a useful discussion of how to optimally allocate a vaccine when it has not been produced in sufficient quantities for all (although he does not go into the issue of people refusing it, which is likely to be a problem). And he considers how innovation in vaccine development can be motivated without reliance on market forces and patenting of what is so clearly a global public good. There is previous discussion he might have drawn on here about the development of urgently needed drugs and treatments that seem unlikely to generate profits for pharmaceutical companies, such as new antibiotics and treatments for tuberculosis.

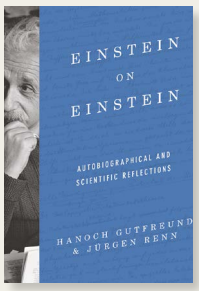
As others have done after previous outbreaks, Gans advocates establishing a pan-national institution with a "set of resources to contain future pandemics and ensure an international, harmonized response". More like the International Monetary Fund than the WHO in his vision, this would focus not just on drugs but also on innovations to enhance protection from infection at work and on public transport. He calls "hundreds of billions of dollars per year to mitigate substantially the risk of global pandemics" a no-brainer, echoing those after the first Ebola outbreak who drew parallels with defence spending.

Here, the book stops short. Realistically, Gans's word was always going to be the first, not the last. But he paints a picture of a post-COVID-19 world that is largely back to normal, with some inconveniences. The truth is that the pandemic throws much more into question. Whatever landscape emerges, it is unlikely to be same as that at the end of 2019.

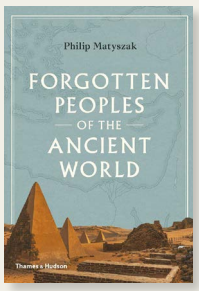
There is a moral case for rethinking inequalities in light of what we have learnt about who is truly essential for society's functioning. Some aspects of neoliberal economic policy are fundamentally in conflict with the needs of a fragile world, with greater risks to come. The behaviour of some leaders has pointed out the dangers of an information economy that has become a 'marketplace for truth'. And economics itself must incorporate the revision of past preconceptions and habits that will be demanded of the rest of us.

Philip Ball is a science writer and author based in London. His latest book is *How to Grow a Human*.
e-mail: p.ball@btinternet.com

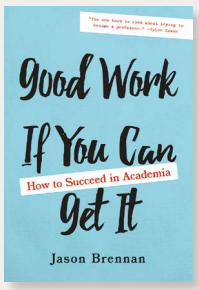
Books in brief



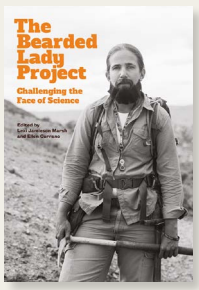
Einstein on Einstein
Hanoch Gutfreund & Jürgen Renn Princeton Univ. Press (2020)
Albert Einstein admitted in his final essay, 'Autobiographical sketch', that fellow physicists opposed his quest to unify the general theory of relativity with quantum mechanics. But he took comfort from philosopher Gotthold Lessing's dictum: "The search for truth is more precious than its possession." The 1955 work appears in English for the first time in this outstanding study of another essay, which Einstein called his "obituary": 1949's 'Autobiographical notes'. Physicist Hanoch Gutfreund and historian Jürgen Renn provide a sparky commentary.



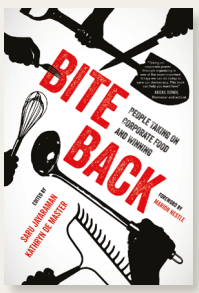
Forgotten Peoples of the Ancient World
Philip Matyszak Thames & Hudson (2020)
Western ideas on antiquity are dominated by Egyptians, Babylonians, Assyrians, Hebrews, Greeks and Romans, with other cultures often reduced to stereotypes. Historian Philip Matyszak asks: were the Philistines philistines and the Vandals vandals? His stimulating encyclopaedia of 40 "forgotten peoples" begins with the Akkadians around 2330 BC and ends with the Hephthalites ('White Huns') in the fifth century AD. Illustrations include a Roman-style Vandal mosaic; far from vilifying Roman culture, the Vandals respected it, say current historians.



Good Work If You Can Get It
Jason Brennan Johns Hopkins Univ. Press (2020)
Economist and strategist Jason Brennan delivers a data-driven, punchily practical guide to succeeding in academia, aimed at PhD students. He knows how success requires narrow professionalism, but he also networks widely. And if a PhD does not yield an academic job, all is far from lost. "Faculty jobs are the nail for which the PhD is the hammer," he concludes in the chapter 'Exit options'. Yet the hammer can be repurposed for diverse non-academic jobs: the US unemployment rate for PhD holders is just 1.7%.



The Bearded Lady Project
Eds Lexi Jamieson Marsh & Ellen Currano Columbia Univ. Press (2020)
"Many of our Bearded Ladies became professional palaeontologists because they did not want to spend every workday inside, in an office, behind a desk," writes palaeobotanist Ellen Currano, co-founder of the Bearded Lady Project with film-maker Lexi Jamieson Marsh, in their photo-biography of a weirdly compelling collaboration. It began six years ago, out of despair at male dominance of their professions. "Maybe I should sport a beard," Currano joked. Dozens of female geoscientists have now posed, artificially hirsute.



Bite Back
Eds Saru Jayaraman & Kathryn De Master Univ. California Press (2020)
In this cleverly titled collection, attorney Saru Jayaraman and rural sociologist Kathryn De Master conclude that corporations control much of our food because of "their unbridled, unregulated power over our democracy". Articles on seeds, labour, hunger and more describe calls to action and collective response, such as mobilization of New York state residents to force a ban on fracking, because of its potential to harm farms. Only direct public confrontation with corporate food elites will succeed, the editors argue. **Andrew Robinson**

© 2020 Springer Nature Limited. All rights reserved.

Obituary

Julian Perry Robinson
(1941–2020)

Chemist and lawyer who shaped international weapons conventions.

In 1981, the US government publicly accused Soviet-backed forces in south-east Asia of waging toxin warfare and violating their legal obligations under the 1925 Geneva Protocol and 1972 Biological Weapons Convention. It alleged that aircraft dispersed 'yellow rain' containing mycotoxins that were "not indigenous to the region". Julian Perry Robinson, working alongside biologist Matthew Meselson at Harvard University in Cambridge, Massachusetts, established that what actually fell was wild-honeybee faeces containing naturally occurring toxins. He died on 22 April, aged 78.

This episode illustrates how Robinson helped to bring rationality into a field in which emotions often run high. His ideas influenced the negotiation and implementation of international law. In a major 1970 report for the World Health Organization, he began to articulate the idea that an ability to respond to natural-disease outbreaks could considerably diminish incentives to use chemical and biological warfare (CBW). He called for the strengthening of disease surveillance and other key areas in public health, and suggested that if there was to be "any chance of success", a clear plan was needed for communicating information with the public. His death from complications of COVID-19 is therefore particularly poignant.

Robinson was born in Jerusalem in November 1941. His interest in CBW began during the final year of his chemistry degree at the University of Oxford, UK. Working under the economist John Jewkes, his dissertation examined how the study of chemical warfare during the Second World War stimulated the synthesis of new types of organic compounds. On graduating, he spent four years with Kilburn & Strode Chartered Patent Agents in London before taking a position at the fledgling Stockholm International Peace Research Institute (SIPRI) in 1968, and becoming its focal point for CBW studies. From the mid-1970s, he was also central to the highly influential CBW Study Groups convened by the Pugwash Conferences on Science and World Affairs, which brought together scientists from east and west to discuss disarmament.

At SIPRI he had a prominent role in a major review of CBW. This resulted in the classic six-volume study *The Problem of Chemical and Biological Warfare* (1971–75). He developed the concept of a cross-cultural 'taboo' on such weapons, and detailed the process of 'assimilation' that leads to their acceptance into existing military organizations. For example, chemical weapons went unused during the Second World War not because of deterrence, but because specialized First World War-era chemical warfare organizations were sidelined by conventional military institutions.



CARLOS SABA

He met Mary Kaldor, his partner of more than 50 years, at SIPRI, and they returned to the United Kingdom in 1971 to join the Science Policy Research Unit at the University of Sussex in Brighton. Robinson spent the rest of his career there establishing the Harvard Sussex Program, a group for research, communication and training in support of informed public policy towards CBW issues that he co-directed with Meselson. There, he spotted another barrier to assimilation: existing technologies constrain emerging ones as they pass through a 'weapons succession process'.

> **"Part of Robinson's influence lay in identifying and seeking to close loopholes."**

Part of Robinson's influence lay in identifying and seeking to close loopholes through which suppressed CBW technology might develop. He provided key inputs into the negotiations for the Biological Weapons Convention and the 1993 Chemical Weapons Convention. He advocated for proper implementation of the 'general purpose criterion' – the concept that malign intentions rather than physical objects should be prohibited – as the main mechanism by which the treaties could avoid being undermined by advances in science and technology.

In the years just after the cold war, he stood firm against what he considered to be a "creeping legitimization" of those chemical and biological weapons not tooled for mass destruction (non-WMD CBW). In 2008, he warned that chemical weapons lent themselves particularly to the new types of conflict experienced in places such as the Balkans, Afghanistan and parts of Africa, suggesting the taboo against CBW could become harder to maintain.

These became the hallmark issues that he encouraged numerous research students to tackle. His mentoring of generations of scholars and policy practitioners means his ideas will continue to shape both treaties for decades.

He officially retired from the University of Sussex in 2007, but continued to work at the Harvard Sussex Program, happiest when receiving visitors who came to discuss ideas and work in the extensive and unique archive on CBW issues that he and Meselson established. One of his last research projects resulted in a series of vignettes from across history that he described as "lessons about CBW" for future generations. He also continued to write detailed chronologies – including one on 'novichoks', after the nerve agents were used in Salisbury, UK, in 2018.

I became a student of Julian's in 1996, and continued to work alongside him at the Harvard Sussex Program until he went into self-isolation. One of our last conversations was about establishing a small study group to consider whether his influential chronology on chemical weapons in Syria might reveal patterns that those guarding against CBW should know about. Always modest, he was hesitant about whether people would spend time reading it. When I suggested some might, he smiled and said "Well now, there's a thought. Let's talk more on the other side of all of this."

Caitriona McLeish is a senior fellow and co-director of the Harvard Sussex Program at the Science Policy Research Unit, University of Sussex, UK.
e-mail: c.a.mcleish@sussex.ac.uk

© 2020 Springer Nature Limited. All rights reserved.

Comment



KEVIN FRAYER/GETTY

A woman in Beijing shows a health QR code on her phone to access a shopping area, as a security guard checks her temperature.

Ten reasons why immunity passports are a bad idea

Natalie Kofler and Françoise Baylis

Restricting movement on the basis of biology threatens freedom, fairness and public health.

Imagine a world where your ability to get a job, housing or a loan depends on passing a blood test. You are confined to your home and locked out of society if you lack certain antibodies.

It has happened before. For most of the nineteenth century, immunity to yellow fever divided people in New Orleans, Louisiana, between the 'acclimated' who had survived yellow fever and the 'unacclimated', who had not had the disease[1]. Lack of immunity dictated whom people could marry, where they could work, and, for those forced into slavery, how much they were worth. Presumed immunity concentrated political and economic power in the hands of the wealthy elite, and was weaponized to justify white supremacy.

Something similar could be our dystopian future if governments introduce 'immunity passports' in efforts to reverse the economic catastrophe of the COVID-19 pandemic. The idea is that such certificates would be issued to those who have recovered and tested positive for antibodies to SARS-CoV-2 — the coronavirus that causes the disease. Authorities would lift restrictions on those who are presumed to have immunity, allowing them to return to work, to socialize and to travel. This idea has so many flaws that it is hard to know where to begin.

On 24 April, the World Health Organization (WHO) cautioned against issuing immunity passports because their accuracy could not be guaranteed. It stated that: "There is currently no evidence that people who have recovered from COVID-19 and have antibodies are protected from a second infection"(see go.nature.com/3cutjqz). Nonetheless, the idea is being floated in the United States, Germany, the United Kingdom and other nations.

China has already introduced virtual health checks, contact tracing and digital QR codes to limit the movement of people. Antibody test results could easily be integrated into this

© 2020 Springer Nature Limited. All rights reserved.

system. And Chile, in a game of semantics, says that it intends to issue 'medical release certificates' with three months' validity to people who have recovered from the disease[2].

In our view, any documentation that limits individual freedoms on the basis of biology risks becoming a platform for restricting human rights, increasing discrimination and threatening – rather than protecting – public health. Here we present ten reasons why immunity passports won't, can't and shouldn't be allowed to work.

Ten points

Four huge practical problems and six ethical objections add up to one very bad idea.

COVID-19 immunity is a mystery. Recent data[3] suggest that a majority of recovered patients produce some antibodies against SARS-CoV-2. But scientists don't know whether everyone produces enough antibodies to guarantee future protection, what a safe level might be or how long immunity might last. Current estimates, based on immune responses to closely related viruses such as those that cause severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS), suggest that recovered individuals could be protected from re-infection for one to two years. But if SARS-CoV-2 immunity instead mimics that seen with the common cold, the protection period could be shorter.

Serological tests are unreliable. Tests to measure SARS-CoV-2 antibodies in the blood can be a valuable tool to assess the prevalence and spread of the virus. But they vary widely in quality and efficacy. This has led the WHO and former US Food and Drug Administration commissioner Scott Gottlieb to caution against their use in assessing individual health or immune status. Several available tests are sufficiently accurate, meaning they are validated to have at least 99% specificity and sensitivity. But preliminary data suggest that the vast majority aren't reliable[4]. Low specificity means the test measures antibodies other than those that are specific to SARS-CoV-2. This causes false positives, leading people to think they are immune when they aren't. Low sensitivity means that the test requires a person to have a high concentration of SARS-CoV-2 antibodies for them to be measured effectively. This causes false negatives in people who have few antibodies, leading to potentially immune individuals being incorrectly labelled as not immune.

The volume of testing needed is unfeasible. Tens to hundreds of millions of serological tests would be needed for a national immunity certification programme. For example, Germany has a population of nearly 84 million people, so would require at least 168 million serological tests to validate every resident's COVID-19 immune status at least twice. Two tests per person are the minimum, because anyone who tested negative might later become infected and would need to be retested to be immune certified. Repeat testing, on no less than an annual basis, would be necessary to ensure ongoing immunity. From June, the German government will receive 5 million serological tests a month from the Swiss firm Roche Pharmaceuticals – a leading supplier of one SARS-CoV-2 serological test that has been approved by regulators. This will allow only 6% of the German population to be tested each month.

Even if immunity passports were limited to health-care workers, the number of tests required could still be unfeasible. The United States, for example, would need more than 16 million such tests. At the time of writing, the US Centers for Disease Control and Prevention and US public-health laboratories have performed more than 12 million diagnostic tests for SARS-CoV-2 (3% of the total US population; see go.nature.com/2wemdd2). Even South Korea, a country with high testing rates, had managed to test only 1.5% of its population by 20 May (see go.nature.com/2aztfvp).

Too few survivors to boost the economy. The proportion of individuals known to have recovered from COVID-19 varies widely in different populations. Reports from hot spots in Germany and the United States suggest some locations could have recovery rates between 14% and 30%. In New York state, for example, where 3,000 people were tested at random in grocery shops and other public locations, 14.9% had antibodies against COVID-19 (see go.nature.com/2waaku9). But these seem to be the exception. In an April press conference, the WHO estimated that only 2–3% of the global population had recovered from the virus.

> **"A cafe can't open and serve customers without risk if only a fraction of its staff are certified as immune."**

Low disease prevalence combined with limited testing capacity, not to mention highly unreliable tests, means that only a small fraction of any population would be certified as free to work. Based on current numbers of confirmed US cases, for example, only 0.43% of the population would be certified. Such percentages are inconsequential for the economy and for safety. A cafe can't open and serve customers without risk if only a fraction of its staff are certified as immune. A shop can't turn a profit if only a minuscule proportion of customers are allowed to enter.

Monitoring erodes privacy. The whole point of immunity passports is to control movement. Thus, any strategy for immunity certification must include a system for identification and monitoring. Paper documentation could be vulnerable to forgery. Electronic documentation integrated into a smartphone app would be more resistant to fraud and more effective for contact tracing, retesting and updates of immune status.

But electronic documents present a more serious risk to privacy[5]. In some Chinese provinces, QR codes on smartphones control entrance into public places on the basis of the individual's COVID-19 health status. However, these apps report more than COVID-19 information – including people's locations, travel history, who they've come into contact with and other health information, ranging from their body temperature to whether they've recently had a cold. Taiwan is also using smartphone apps with alert systems that are directly linked to police departments. The United Kingdom, United States and many other countries are testing various app options. Yet there's no guarantee that the apps will recede when COVID-19 does. China has announced that elements of its QR-code tracking system are likely to remain in place after the pandemic ends.

Marginalized groups will face more scrutiny. With increased monitoring comes increased policing, and with it higher risks of profiling and potential harms to racial, sexual, religious or other minority groups. During the pandemic, China has been accused of racially profiling residents by forcing all African nationals to be tested for the virus. In other parts of the world, people from Asia have faced spikes in racialized prejudice.

Before this pandemic, stop-and-frisk laws in the United States already disproportionately affected people of colour. In 2019, 88% of people who were stopped and searched in New York City were African American or Latin American (go.nature.com/2jntjym). And during the pandemic, policing continues to target people from minority groups. Between mid-March and the start of May in Brooklyn, New York, 35 of the 40 people arrested for violating physical distancing laws were black[6].

These numbers are deeply concerning, but would be even more so if monitoring and policing for COVID-19 immunity were to be used for ulterior motives. For example, 'digital incarceration' has already increased in countries such as the United States, Brazil and Iran, where individuals have been released from prison to minimize the spread of COVID-19 and then monitored using digital ankle bracelets. In the United States, where people of colour are racially segregated by neighbourhood and disproportionately incarcerated, digital incarceration could be used to monitor large segments of certain communities. The risk would be even higher if digital monitoring were to be linked to immigration status.

© 2020 Springer Nature Limited. All rights reserved.



LAETITIA VANCON/NYT/REDUX/EYEVINE

Health-care workers in Munich, Germany, take blood to test for antibodies to SARS-CoV-2.

Unfair access. With a shortage of testing, many will not have access. Experience so far suggests that the wealthy and powerful are more likely to obtain a test than the poor and vulnerable. In tiered health-care systems, these inequities are felt even more acutely. In early March, for example, when professional sports teams, technology executives and film celebrities were getting tested, dozens of US states were conducting fewer than 20 tests per day (see https://covidtracking.com/data). The very people who need to get back to work most urgently — workers who need to keep a roof over their head and food on the table — are likely to struggle to get an antibody test. Testing children before they return to school could be a low priority, as would testing retired older people and those who face physical, mental-health or cognitive challenges.

Societal stratification. Labelling people on the basis of their COVID-19 status would create a new measure by which to divide the 'haves' and the 'have-nots' — the immunoprivileged and the immunodeprived. Such labelling is particularly concerning in the absence of a free, universally available vaccine. If a vaccine becomes available, then people could choose to opt in and gain immune certification. Without one, stratification would depend on luck, money and personal circumstances. Restricting work, concerts, museums, religious services, restaurants, political polling sites and even health-care centres to COVID-19 survivors would harm and disenfranchise a majority of the population.

Social and financial inequities would be amplified. For example, employers wanting to avoid workers who are at risk of becoming unwell might privilege current employees who have had the disease, and preferentially hire those with 'confirmed' immunity.

Immunity passports could also fuel divisions between nations. Individuals from countries that are unable or unwilling to implement immunity passport programmes could be barred from travelling to countries that stipulate them. Already people with HIV are subjected to restrictions on entering, living and working in countries with laws that impinge on the rights of those from sexual and gender minorities — such as Russia, Egypt and Singapore.

New forms of discrimination. Platforms for SARS-CoV-2 immune certification could easily be expanded to include other forms of personal health data, such as mental-health records and genetic-test results. The immunity passports of today could become the all-encompassing biological passports of tomorrow. These would introduce a new risk for discrimination if employers, insurance companies, law-enforcement officers and others could access private health information for their own benefit. Such concerns have been catalogued over the past few years in debates about who should have access to genetic information, as demand rises from clinicians, researchers, insurers, employers and law enforcers, for example[7].

Threats to public health. Immunity passports could create perverse incentives. If access to certain social and economic liberties is given only to people who have recovered from COVID-19, then immunity passports could incentivize healthy, non-immune individuals to wilfully seek out infection — putting themselves and others at risk[8]. Economic hardship could amplify the incentive if an immunity passport is the only way to a pay cheque. Individuals might obtain documents illicitly, through bribery, transfer between individuals or forgery. These could create further health threats, because people claiming immunity could continue to spread the virus. Crises tend to foster nefarious trade, as happened during the Second World War when food rations in Britain caused the emergence of a robust underground exchange system.

Next steps

Strategies that focus on the individual — using conceptions of ethics rooted in libertarianism — contradict the mission of public health[9]. They distract attention from actions that benefit all, such as funding international collaborations, practising effective public-health measures and redressing income inequity. In North America (and elsewhere), because of structural inequities, people of colour are dying from COVID-19 at much higher rates than are white people, and the virus is disproportionately affecting those who live in First Nations territories. Success depends on solidarity, a genuine appreciation that we are all in this together. An ethic premised on individual autonomy is grossly inappropriate during a public-health crisis; the overarching aim must be to promote the common good.

Instead of immunity passports, we contend that governments and businesses should invest available time, talent and money in two things.

First is the tried and true formula of pandemic damage limitation — test, trace and isolate — that has worked well from Singapore and New Zealand to Guernsey and Hanoi. Health status, personal data and location must be anonymized. Apps that empower individuals to make safe choices about their own movements should be prioritized.

Second is the development, production and global distribution of a vaccine for SARS-CoV-2. If universal, timely, free access to a vaccination becomes possible, then it could be ethically permissible to require vaccine certification for participation in certain activities. But if access to a vaccine is limited in any way, then some of the inequities we highlight could still apply, as the literature on uptake of other vaccines attests[10].

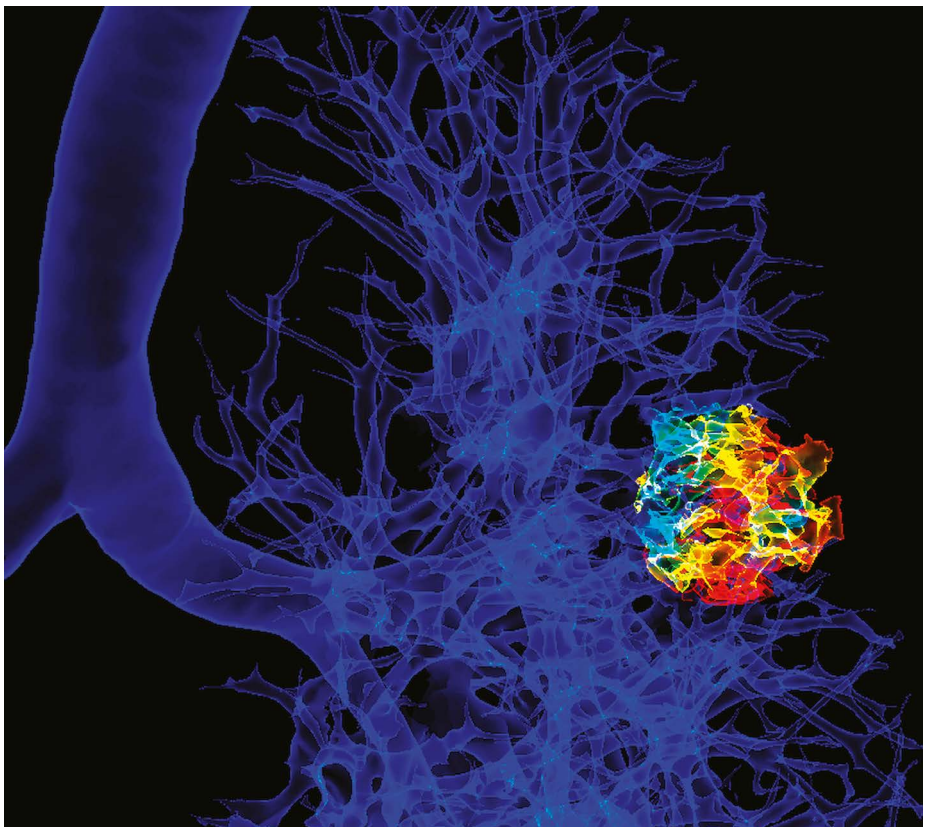
Threats to freedom, fairness and public health are inherent to any platform that is designed to segregate society on the basis of biological data. All policies and practices must be guided by a commitment to social justice.

The authors

Natalie Kofler is founder of the global initiative Editing Nature and adviser for the Scientific Citizenship Initiative, Harvard Medical School, Boston, Massachusetts, USA. **Françoise Baylis** is a professor of philosophy with a specialization in bioethics at Dalhousie University, Halifax, Nova Scotia, Canada.
e-mail: natalie.kofler@hms.harvard.edu

1. Olivarius, K. *Am. Hist. Rev.* **124**, 425–455 (2019).
2. Fraser, B. *Lancet* **395**, 1473 (2020).
3. Wajnberg, A. *et al.* Preprint at medRxiv https://doi.org/10.1101/2020.04.30.20085613 (2020).
4. Whitman, J. D. *et al.* Preprint at medRxiv https://doi.org/10.1101/2020.04.25.20074856 (2020).
5. Ienca, M. & Vayena, E. *Nature Med.* **26**, 463–464 (2020).
6. Southall, A. *The New York Times* (7 May 2020).
7. Clayton, E. W., Evans, B. J., Hazel, J. W. & Rothstein, M. A. *J. Law Biosci.* **6**, 1–36 (2019).
8. Phelan, A. L. *Lancet* https://doi.org/10.1016/S0140-6736(20)31034-5 (2020).
9. Baylis, F., Kenny, N. P. & Sherwin, S. *Public Health Ethics* **1**, 196–209 (2008).
10. Bocquier, A., Ward, J., Raude, J., Peretti-Watel, P. & Verger, P. *Expert Rev. Vaccines* **16**, 1107–1118 (2017).

© 2020 Springer Nature Limited. All rights reserved.



A 3D image of a lung tumour (multicoloured), obtained using a computed tomography scan.

K. H. FUNG/SPL

Take lessons from cancer evolution to the clinic

Charles Swanton

The first long-term study of how lung cancer evolves is revealing that therapies targeting multiple proteins in tumour cells could help to outpace the disease.

How non-small-cell lung cancer evolves in individual patients is being studied in a project called TRACERx (Tracking Cancer Evolution through Therapy), the first such large-scale, longitudinal study to do so. Insights from the project so far are compiled this week in a collection in *Nature* (www.nature.com/collections/haffgaicaf). Started in 2014, the project aims to follow 840 people being treated at 14 UK National Health Service hospitals, from diagnosis through to cure or relapse and, tragically, death (see go.nature.com/2vkxwdy). Some tantalizing results have begun to emerge. These help to explain why early diagnosis is so crucial for effective treatment, as clinicians have noted for decades.

Most of researchers' understanding of cancer evolution has come from observing cellular changes in biopsies of tumours or alterations in DNA from tumour cells. Such samples are typically taken for diagnosis, and provide only snapshots of a complex process. By contrast, TRACERx aims to monitor this process as it unfolds. The study compares molecular changes in the responses of people's tumour cells and in immune cells, across multiple tumour regions and as the tumour progresses.

Scientists in the TRACERx project, for which I am chief investigator, hope that clinical practice might soon benefit from the insights emerging, combined with those from previous studies on cancer evolution. As well as explaining a person's immune response against their cancer, the study could guide treatments for targeting the many genetic mutations that accumulate in late-stage cancers that are difficult to treat. More than 80% of people who are diagnosed with stage-four lung cancer do not survive for more than five years.

Project approach

Before the COVID-19 pandemic, the TRACERx project had recruited 760 people with early-stage lung cancer. After a person is diagnosed with a primary lung tumour, it is surgically removed and the cells are analysed to reconstruct the tumour's evolutionary history. Each individual receives a computed tomography (CT) scan every year for five years to check whether their cancer has returned. If there is no sign of relapse, they are discharged and deemed to have been cured. People with later-stage tumours (stages 2 and 3) are offered chemotherapy following surgery to improve the chance of remission or cure.

Analysis of tumours from the first 100 people enrolled in the study has revealed many genomic changes. These include chromosome deletions and duplications, and even the doubling of whole genomes in nearly three-quarters of tumours — a feature of many cancers[1]. Point mutations in DNA, arising from single changes in the genome sequence, were also prevalent. These occurred as a result of tobacco exposure and the activity of enzymes called cytidine deaminases, which normally deactivate invading viruses as part of the immune response.

Another finding is that whole-genome doubling often occurs early on in those with lung cancer who have a history of smoking[1]. This doubling seems to protect the genes needed for the tumour's survival in the face of the excessive mutations and chromosomal losses that occur in its genome as it develops[2]. Strikingly, mutations induced by smoking tend to dominate the 'trunk' of the tumour's evolutionary tree. These are known as founder or truncal mutations, and are present in every tumour cell. For the most common type of non-small-cell lung cancer — adenocarcinomas that form in mucus-secreting glands — the number of smoking-related mutations in the trunk correlates with the number of cigarettes that the person has smoked. As the cancer advances, cytidine deaminases cause haphazard mutations to accumulate in some cells; we refer to these as branched mutations[2].

With a view to devising immunotherapy[3], we also investigated the DNA sequences of receptors on T cells, a type of white blood cell that fights off infection and emerging cancers. We were surprised to discover that the sequences of T-cell receptors evolved in parallel with the tumour. (The tumour eventually adapts to its immune environment and avoids destruction[3,4]). One possible explanation for the tumour's ability to evade the immune system is the chromosomal instability of tumour

© 2020 Springer Nature Limited. All rights reserved.

cells. This causes the tumour cells to lose immune-recognition molecules (called HLA, for human leukocyte antigen) on their surface, preventing T cells from homing in on the tumour[5,6].

This unstoppable genetic diversification can outwit cancer drugs. So our aim is to boost the immune response to target the mutations present in every tumour cell and prevent the development of drug resistance.

Steer for clinicians

How might the discoveries from TRACERx and previous studies on cancer evolution guide clinical practice?

The overriding message is that it will take multiple approaches to outpace cancers that have sophisticated evolutionary mechanisms, such as lung cancer. Each approach will have to focus on different stages of the disease to improve outcomes.

For example, lung-cancer screening should be rolled out nationwide: earlier diagnosis improves prognosis. In the United Kingdom, lung-cancer screening so far occurs in only a few regional centres. On the basis of results from a clinical trial in the Netherlands, I calculate that screening could prevent approximately 3,000–5,000 deaths from lung cancer in the United Kingdom every year (unpublished work).

In the Netherlands trial, 13,195 male and 2,594 female former or current smokers aged 50 to 74 were split at random into two groups. One group received CT screening when they were diagnosed and at years 1, 3 and 5.5; the other received no screening. After 10 years of follow up, there were 2.5 deaths per 1,000 person years in the screened group[7]. (Put another way, this means there would be, on average, 2.5 deaths from lung cancer in the screened group if 1,000 people in the Netherlands were observed for one year.) There were 3.3 deaths per 1,000 person-years in the control group[7]. This equates to a 24% reduction in the chance of death from lung cancer if CT screening is done.

Treatments that are tailored to an individual's needs can also improve their prognosis. For example, some people benefit from chemotherapy after surgery to destroy any residual tumour cells following removal of the primary tumour. For every 100 people who are given chemotherapy after surgery, between 5 and 15 extra patients will be alive (depending on tumour stage) after 5 years, compared with those who do not receive chemotherapy[8]. Currently, doctors cannot predict who those individuals might be.

The TRACERx study has come up with a sensitive method to predict who is most likely to relapse following surgery — by identifying truncal DNA mutations that are common to all lung-cancer cells in a patient's tumour. This tumour DNA appears in the blood as circulating free DNA after surgery — a signal of minimal residual disease. Our patented sequencing approach seeks residual tumour DNA harbouring these truncal mutations[9] and is now ready for clinical trials (see competing-interests declaration at go.nature.com/2w214yu). These trials could test the hypothesis that post-operative chemotherapy is beneficial only to patients who have tumour DNA in their blood after surgery, and therefore have minimal residual disease.

The TRACERx study also offers insight into how immunotherapies for lung cancer could be refined. Currently, the malignant cells that some cancer therapies target carry one mutant protein on their surface that drives tumour expansion, such as a growth-factor receptor that is jammed in the 'on' position. In most people, cancer cells become resistant to such therapies after 18 months, through acquired resistance mutations in the receptor or the downstream pathway. Knowing that the immune system adapts rapidly in tandem with the tumour, clinicians could now target not just one but many mutant cancer-cell proteins that are present in every tumour cell, and which occur early in tumour evolution (truncal mutations). This might decrease the chances of resistance developing, and could have the advantage of not harming normal tissue. It might also provide a way to keep abreast of complex, fast-evolving cancers.

Clinical trials are now exploring how to exploit these truncal mutations, by extracting the T cells that recognize them from a patient with cancer, culturing the cells in the laboratory and transferring them back into the patient to amplify his or her immune response against the truncal mutations. Other UK researchers hope to assess the effectiveness of these cells in limiting tumour growth and spread (see competing-interests declaration).

Future challenges

Despite the promise set out here, many scientific and clinical challenges remain when it comes to cancer evolution.

TRACERx and other studies have revealed the complex interactions between malignant tumour cells and normal cells in the surrounding environment[10]. They can even come to depend on each other for survival and growth. Our understanding of these dynamics is still rudimentary. To assess them over space and time, a patient's tumour would need to be sampled frequently, which is unethical and not possible. The TRACERx project includes a national autopsy study called PEACE that helps to resolve this problem and sample more deeply (see go.nature.com/2kutvyr).

Scientists on the PEACE study are analysing the tumour material of 20 patients from TRACERx who gave consent for their tissues to be used in research after their death. Work in progress is beginning to reveal how cancer cells spread from the primary tumour to distant sites in the body. TRACERx is also illuminating how cells that are shed from the tumour at surgery (circulating tumour cells) reflect the future site of spread months later[11]. Developments in single-cell sequencing technologies will enhance our understanding of this process.

Animal models that are used to probe and manipulate cancer evolution can mimic certain key features, but they do not completely mirror human disease. Mouse lung-cancer cells, for instance, carry just 1% of the mutations found in cancer cells of people who smoke or have smoked[12]. There are no mouse lung-cancer models for tumour mutations that arise from the enhanced activity of cytidine deaminase. Moreover, in mice, the immune mechanisms that evade cancer cannot yet be effectively manipulated in parallel with an evolving human tumour that has many mutations. Researchers need to develop mouse models that better mimic how human lung cancers evolve and evade the immune system. Gene-editing tools such as CRISPR–Cas9 could help here. Such models would be valuable for drug development and preclinical testing.

Approaches are beginning to emerge to image 'normal' cells in the microenvironment of a growing tumour; some of these techniques are being used by the TRACERx consortium. Advances in sequencing nucleic acids and proteins from single cells — preferably performed *in situ* without disrupting the tumour architecture — should help to decipher details of the myriad steps in cancer evolution[13].

Multiple efforts worldwide provide hope for the future of lung-cancer treatment. These include programmes to minimize tobacco use, more CT screening to detect early lung cancers, clinical trials focusing on minimal residual disease, and developments in immunotherapy for advanced tumours. This combination holds much promise for improving the survival and quality of life for patients with lung cancer.

The author

Charles Swanton is chief investigator of the Cancer Research UK TRACERx clinical study, and group leader of the Cancer Evolution and Genome Instability Laboratory at the Francis Crick Institute and UCL Cancer Institute, London.
e-mail: charles.swanton@crick.ac.uk

1. Jamal-Hanjani, M. *et al. N. Engl. J. Med.* **376**, 2109–2121 (2017).
2. López, S. *et al. Nature Genet.* **52**, 283–293 (2020).
3. Joshi, K. *et al. Nature Med.* **25**, 1549–1559 (2019).
4. Ghorani, E. *et al. Nature Cancer* (in the press).
5. McGranahan, N. *et al. Cell* **171**, 1259–1271 (2017).
6. Rosenthal, R. *et al. Nature* **567**, 479–485 (2019).
7. de Koning, H. J. *et al. N. Engl. J. Med.* **382**, 503–513 (2020).
8. Pignon, J.-P. *et al. J. Clin. Oncol.* **26**, 3552–3559 (2008).
9. Abbosh, C. *et al. Nature* **545**, 446–451 (2017).
10. AbduJabbar, K. *et al. Nature Med.* (in the press).
11. Chemi, F. *et al. Nature Med.* **25**, 1534–1539 (2019).
12. McFadden, D. G. *et al. Proc. Natl Acad. Sci. USA* **113**, E6409–E6417 (2016).
13. Rozenblatt-Rosen, O. *et al. Nature* **550**, 451–453 (2017).

The author declares competing financial interests; see go.nature.com/2w214yu.

© 2020 Springer Nature Limited. All rights reserved.

Correspondence

Rwanda's success in tackling COVID-19

Rwanda's strong health-care system and strictly coordinated prevention measures against COVID-19 have helped the country to record zero deaths from the disease so far. As the pandemic threatens to gather momentum in Africa, other governments there could benefit from lessons we have learnt.

Rwanda implemented full lockdown a week after its first case was reported in mid-March. A week later, it set up a contact-tracing system and implemented testing for all staff policing borders, as well as those working in public spaces such as banks and bars. By the end of April, 29,395 citizens had been tested for COVID-19 (prevalence was 0.7%). The nation's community health network has enabled the government — with help from the private sector — to identify populations in need of extra support.

Africa has so far recorded relatively few cases and deaths compared with other continents (https://covid19.who.int). Strict prevention measures that are coordinated across countries could keep it that way. Regional bodies such as the East African Community should agree guidelines for full lockdown, backed by surveillance and a supranational testing laboratory, and follow up with population-impact surveys for mental health and COVID-19 serological status.

Jeanine Condo University of Rwanda, Kigali, Rwanda.
jennycondo@gmail.com

Jean Paul Uwizihiwe Centre for Global Health, Aarhus University, Denmark.

Sabin Nsanzimana Rwanda Biomedical Centre, Kigali, Rwanda.

Developing world: boost modelling

Computational models of the likely spread of the SARS-CoV-2 coronavirus have been instrumental in guiding governments' strategies to limit disease transmission and control the current public-health crisis. In developing countries, where the pandemic is potentially at its most dangerous and costly, we call for governments to work with academic institutions to build and sustain modelling capacity.

Models are not silver bullets for fixing the ills of developing countries. Nevertheless, partnerships with international academic modelling communities (see, for example, F. Squazzoni *et al. J. Artif. Soc. Soc. Simul.* **23**, 10; 2020) and the participation of stakeholders and experts from different disciplines could help to build useful models.

Such models would combine knowledge of computational techniques with local contextual knowledge of social processes. They would enable policymakers to distil choices from uncertainties, particularly when stakes are high and resources limited. Once set up, they could be used in times of both crisis and calm.

Kaveri Iychettira, **Afreen Siddiqi** Belfer Center for Science and International Affairs, Harvard Kennedy School, Cambridge, Massachusetts, USA.
kaveri_iychettira@hks.harvard.edu

Spotlight on figures for COVID-19

Given the importance of accurate reporting of COVID-19 cases and deaths to strategies for virus control, we screened countries' daily records for possible misreporting by applying Benford's law. This can pick up unreliable numbers resulting from error, oversight or manipulation, for instance — although it cannot distinguish their possible causes.

Benford's law predicts the relative frequency distribution of first digits in real-world number sets (see M. Sambridge *et al. Geophys. Res. Lett.* **37**, L22301, 2010). Anomalies have exposed financial fraud, for example (see M. J. Nigrini *J. Am. Tax Assoc.* **18**, 72–91; 1996). We tested data reported by 51 countries from 16 January until 9 April 2020, when case numbers were still on the rise. Once these start to level out, as was the case for China and South Korea over that period, Benford's law can no longer be applied.

We found that records of cumulative infections and deaths from the United States, Japan, Indonesia and most European nations adhered well to the law (see go.nature.com/2kqtut2) and therefore are consistent with accurate reporting. Figures from a few of the countries analysed reveal anomalies. These could be explained by limited data sets or by adjustments to avoid headline-grabbing numbers of deaths in the hundreds or thousands.

Malcolm Sambridge Australian National University, Canberra, Australia.
malcolm.sambridge@anu.edu.au

Andrew Jackson Swiss Federal Institute of Technology, Zurich, Switzerland.

Can public trust coronavirus apps?

On 6 April, we approached the Belgian government with concerns about the improper use of contact-tracing smartphone apps for controlling pandemics. These concerns were in line with those you discuss (*Nature* **580**, 563; 2020). On 17 April, we drew its attention to other issues relating to lockdown exit strategies.

We argued that contact-tracing apps could complicate rather than facilitate lockdown exit (see go.nature.com/36ebfmq). For example, receiving (or not) a warning through the app might elicit a false sense of security, or drive demand for testing that might not be available. And there is more at stake than the government's public-health-efforts and investment: an app's success also depends on personal, public and social trust.

Governments need to engage stakeholders to co-design the app so that it aligns with local culture and connects with vulnerable populations. They also need to use proper information campaigns and human follow-up after issuing app warnings, and to ensure that the media accurately relay what the apps can and cannot deliver.

For now, the Belgian government has paused its implementation of contact-tracing apps (see go.nature.com/2zinmbb). If they pursue the project, we hope it will incorporate the necessary caution and guidance for citizens.

Anne-Mieke Vandamme*, **ToTran Nguyen*** Institute for the Future, KU Leuven, Leuven, Belgium.
annemie.vandamme@kuleuven.be
*On behalf of the Coronavirus Pandemic Preparedness team (see go.nature.com/2xemdt6).

© 2020 Springer Nature Limited. All rights reserved.

News & views

Human genetics

Ability to understand genomes scales up

Deanna M. Church

A massive genome-sequencing and analysis effort has produced the most comprehensive sets of data and tools for understanding human genetic variation so far. The resource will be invaluable to biologists of every stripe. **See p.434, p.444, p.452 & p.459**

What do the differences between each person's genetic code mean for their individual development and health? Several factors have hampered researchers' ability to answer this question. First, understanding genetic variation requires analysing huge numbers of sequences, because we carry many rare variants[1]. Most of these have no effect, with just a few causing genetic diseases. Second, most of our understanding of genetic variation has come from studying single nucleotide variants (SNVs), but structural variants – more than 50 nucleotides long – can have a larger impact on physiological traits, and are major contributors to disease[2]. Third, we lack an understanding of variation outside protein-coding sequences. In four papers in *Nature*, the genome aggregation database (gnomAD) consortium[3–6] sets out to address these gaps in knowledge.

The gnomAD project is the successor to the game-changing exome aggregation consortium (ExAC)[7] project, which catalogued genetic variation in the protein-coding parts of the genome, called exomes, from more than 60,000 people (Fig. 1). ExAC set a new standard for harmonized analysis – bringing in data from diverse projects for reanalysis in a common pipeline – and for data sharing. The ExAC data were available to scientists well before the project's publication in 2016, and it has had a profound impact on how researchers, physicians and genetic counsellors interpret the genomes of people with genetic diseases.

In the first of the current papers, Karczewski *et al.*[3] describe the gnomAD consortium's collection of 125,748 exomes and 15,708 whole genomes. The move to sequencing whole genomes is especially exciting, because analysis of non-coding sequences provides information about both structural variation and variation in DNA sequences that regulate gene expression – described in the companion papers. The gnomAD resource includes sequences from diverse populations, including individuals from Asia and Africa. However, as the authors note, representation from more diverse populations is still needed to obtain the full spectrum of human variation and to capture more population-specific variation.

> **"The impact that the project will have on science goes well beyond the current collection."**

Karczewski *et al.* went on to analyse the protein-coding variants in their data set. They built on a metric developed by the ExAC group to assess whether a gene can 'tolerate' variants that are predicted to prevent the normal functions of the protein it encodes – that is, whether these predicted loss-of-function (pLoF) variants have little to no effect on physiology, or cause serious health issues or death. This type of analysis is useful because genes that are intolerant to LoF might be essential for life, or their mutation could cause genetic diseases.

The ExAC metric measures how many pLoF variants are observed in a gene across a population, compared with how many are expected given the rate at which mutations arise in genomes throughout evolution. However, because pLoF variants are so rare, 60,000 exomes was not enough to definitively say whether all the genes studied – particularly small genes – are intolerant to pLoF. The data were therefore expressed as the probability that a given gene would tolerate pLoF.

By contrast, the increased cohort size in gnomAD allows for a more-direct measure of gene tolerance to LoF. Karczewski *et al.* binned genes into ten groups according to the frequency of pLoF variants they contained compared with those expected, producing a

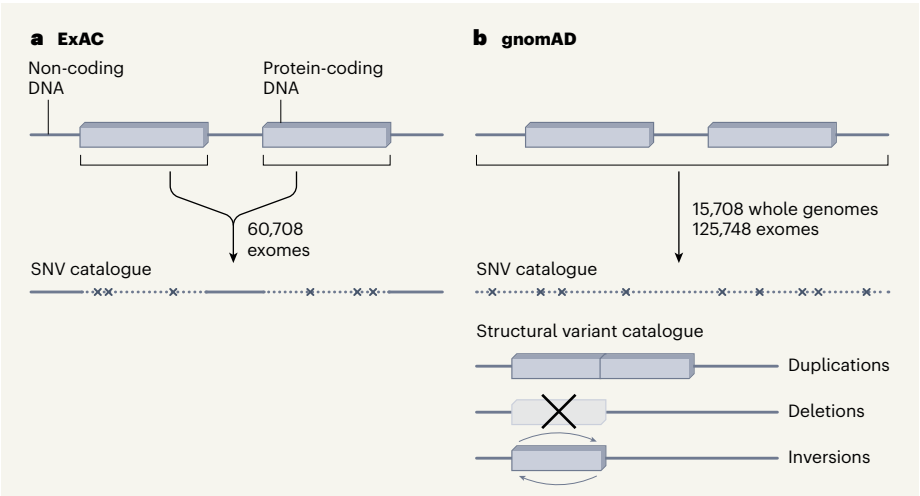


Figure 1 | Cataloguing genetic variation in humans. a, In 2016, the exome aggregation consortium (ExAC) catalogued sites in the protein-coding sequences (the exomes) of the human genome at which single nucleotides could vary between individuals[7]. The database was formed of exomes from 60,708 people. **b**, Its successor, the genome aggregation database (gnomAD) includes 15,708 whole-genome sequences, in addition to 125,748 exomes[3–6]. The consortium catalogued not only single-nucleotide variants (SNVs) across the whole genome, but also more-complex structural variants, which span 50 nucleotides or more. These can include deletions, inversions or duplications of DNA.

© 2020 Springer Nature Limited. All rights reserved.

spectrum of pLoF tolerance. The larger sample size means that gene length is less of a problem in the gnomAD analysis, but even so, the authors could not definitively assess pLoF frequency in the 30% of genes that were expected to have few pLoF variants.

Despite this limitation, the group use their approach to gain fresh insights into the genetics of disease. For example, they found rare variants in genes that do not tolerate LoF more often in people who have an intellectual disability or autism spectrum disorder than in people who do not. These data might help researchers to understand the complex genetic structure that underlies these traits.

In the second paper of the collection, Cummings *et al.*[4] investigated why genes that seem intolerant to pLoF can sometimes carry these variants with apparently little consequence. Genes can be transcribed in different ways, with some protein-coding regions (exons) expressed only in a limited fashion. Cummings and colleagues demonstrated that, when an individual carries a pLoF variant in an 'intolerant' gene, the variant is often in an exon that shows this restricted expression, thus limiting its effect.

In the third paper, Minikel *et al.*[5] assessed how the pLoF database might improve our ability to identify genetic targets for drugs. The identification of individuals who carry two pLoF variants in a given gene is desirable in drug discovery — if these individuals also exhibit a change in a particular trait, it provides evidence that the gene could be a good drug target[8]. The group showed that there are still many errors when identifying pLoF variants; that quality control is needed when identifying these variants; and that instances of an individual carrying two pLoF variants in the same gene are sufficiently rare that we will need cohorts roughly 1,000 times bigger than gnomAD to gather definitive evidence of their existence in most genes.

One of the most exciting aspects of the gnomAD project is the production of a catalogue of structural variants, described in the final paper by Collins and co-workers[6]. There have been excellent efforts at cataloguing structural variants using long-read sequencing technology[9]. However, sample sizes have been small, owing to the expense and lack of standardized analysis pipelines for this approach — although I expect this situation to improve in the near future. By contrast, identifying structural variants in short-read sequences is technically challenging, because the variants are often larger than a typical short sequence read, and they can arise through a variety of mutational mechanisms, resulting in many variant types (duplication, deletion or inversion of DNA, for instance) that each leave different footprints in the genome. This has led to the development of many tools for identifying structural variants from short reads, but no 'standard' pipeline.

Collins *et al.* sought to remedy this problem by creating a pipeline that allows for harmonized analysis over thousands of genomes; this could become the industry standard for structural-variant detection from short-read sequences on a population scale. The authors generated a catalogue of more than 300,000 high-quality structural variants — more than twice as many as previous analyses. They then began to assess the contribution of structural variants to physiological traits. This analysis revealed some evidence for natural selection against structural variants in non-coding sequences that control gene expression. Unsurprisingly, selection against structural variants was stronger in protein-coding regions. This suggests that more variation is tolerated in non-coding than in coding regions, and that even-larger cohorts (or other approaches) will be needed to begin to robustly dissect non-coding variation. The authors also found that structural variants account for roughly one-quarter of protein-truncating events.

The routine analysis of structural variants, integrated with analysis of SNVs and gene expression, will be crucial for interpreting individual genomes. Collins *et al.* have taken an important step in this direction, and the gnomAD resource provides tools for others to continue on this path.

An interesting, recurring theme in these papers is that — despite the size of the cohort — we still lack the numbers required for many analyses. The sequencing of ever-larger cohorts should no doubt continue. However, this approach alone will not enable us to fully understand the relationships between human genetics and traits at both cellular and organismal levels. We need scalable approaches to program genetic variation into human cells, and well-characterized cellular traits that can be monitored to allow us to directly interrogate the physiological impact of this variation. Such interventional biology will substantially augment population genetics and accelerate our understanding of human biology.

The gnomAD consortium has already made its data publicly available. The impact that the project will have on science goes well beyond the current collection, which includes not only the papers in this issue, but several published in *Nature*'s sister journals (go.nature.com/2zgfxr2). The gnomAD resource, like ExAC before it, will change how we interpret individual genomes. The consortium's work has revealed how much information about human variation we had been missing and has provided tools that help us to better understand the genome at both the population and individual level. I can't wait to see what comes next.

Deanna M. Church is at Inscripta, Boulder, Colorado 80301, USA.
e-mail: deanna.church@inscripta.com

1. Auton, A. *et al. Nature* **526**, 68–74 (2015).
2. Chiang, C. *et al. Nature Genet.* **49**, 692–699 (2017).
3. Karczewski, K. J. *et al. Nature* **581**, 434–443 (2020).
4. Cummings, B. B. *et al. Nature* **581**, 452–458 (2020).
5. Minikel, E. V. *et al. Nature* **581**, 459–464 (2020).
6. Collins, R. L. *et al. Nature* **581**, 444–451 (2020).
7. Lek, M. *et al. Nature* **536**, 285–291 (2016).
8. Plenge, R. M., Scolnick, E. M. & Altshuler, D. *Nature Rev. Drug Discov.* **12**, 581–594 (2013).
9. Chaisson, M. J. P. *et al. Nature Commun.* **10**, 1784 (2019).

The author declares competing financial interests: see go.nature.com/3eklyoi for details.

Materials science

A model of perfection for light-activated catalysts

Simone Pokrant

Efforts to make hydrogen from water directly using sunlight have been hampered by the inefficiency of the catalysts that promote the process. A model system demonstrates that almost perfectly efficient catalysts can be made. **See p.411**

Since the emergence of Greta Thunberg's 'Fridays For Future' movement in August 2018, the need to prevent climate change and to find 'green' alternatives to fossil fuels have become topics of broad public interest. But although public awareness has advanced rapidly, progress in the search for cost-effective technological solutions has not. One promising sustainable energy carrier is hydrogen, if it can be produced using renewable energy sources — hydrogen is a green fuel, because its combustion produces only pure water. On page 411, Takata *et al.*[1] report a breakthrough in catalyst design that might accelerate the development of large-scale processes for making hydrogen from water using sunlight.

The largest potential source of renewable energy is the Sun[2]: about 0.02% of the solar

© 2020 Springer Nature Limited. All rights reserved.

energy absorbed by Earth's surface annually would be enough to cover current global energy consumption. Many approaches for converting solar energy into the chemical energy stored in hydrogen are therefore being investigated, using 'water-splitting' reactions in which water is broken down into hydrogen and oxygen. Some of these approaches are already being tested in pilot facilities – such as solar-power-to-gas units, in which electricity produced by solar cells is used to split water through electrolysis[3]. Hydrogen produced in this way could be used for the long-term storage of solar energy, or as fuel for vehicles. However, solar-to-gas conversion processes are not yet economically viable.

A study of the technical and economic feasibility of solar-energy production[4] has shown that systems based on light-activated catalysts (photocatalysts) are attractive alternative options for water splitting. In these systems, photocatalytic semiconductor particles are suspended in a bed filled with an aqueous electrolyte; when sunlight shines on the suspension, hydrogen and oxygen gases are produced. The technical simplicity of this approach should enable economically competitive hydrogen production, if the photocatalyst can convert solar energy to hydrogen with a minimum efficiency of about 10%.

However, the conversion efficiencies of photocatalytic semiconductors are typically much lower than 10%. This is because the photocatalytic process is highly complex and requires the semiconductor particles to have a combination of several properties. They must: absorb light; generate and separate electron–hole pairs (holes are positively charged quasiparticles produced when photons knock electrons out of an atomic lattice); enable holes and electrons to travel to the particle–water interface; and catalyse the production of hydrogen and oxygen from water (reactions that require electrons and holes, respectively). Side processes that can occur at each step can lower the overall conversion efficiency. Materials scientists are therefore trying to design photocatalysts that minimize such efficiency losses.

A key measure of the effectiveness of a photocatalyst is the fraction of absorbed photons that it can use to produce hydrogen, a quantity called the internal quantum efficiency (IQE). A perfect photocatalyst that converts all of the absorbed photons to hydrogen would have an IQE of 1 (or 100%). However, IQE cannot be determined from experiments.

A related quantity that can be experimentally determined for a reaction is the external quantum efficiency (EQE): the fraction of photons illuminating the reaction vessel that the photocatalyst can use to produce hydrogen. This value is always lower than the IQE, because an unknown portion of the illuminating photons will not be absorbed by the photocatalyst, but will instead be lost to other processes, such as scattering. If similar photocatalyst-particle suspensions are investigated using the same experimental set-up, ensuring that the same fraction of light is absorbed, then EQE can be used as an indirect measure of IQE. But EQEs determined using different set-ups cannot be used as a way of comparing IQEs of photocatalytic systems, because the relationship between EQE and IQE is different for each set-up – therefore making it difficult for different research groups to compare results.

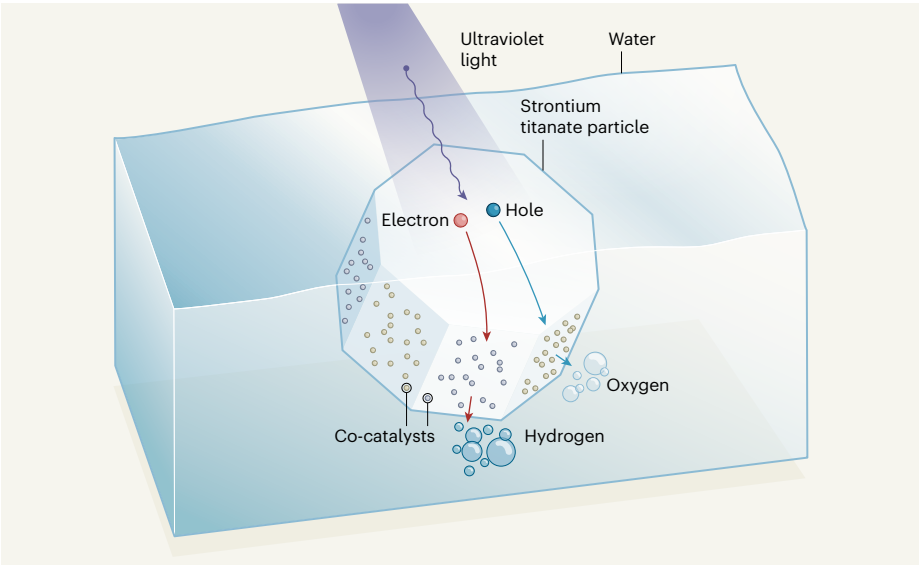


Figure 1 | A photocatalytic particle engineered for water splitting. Particles of light-activated catalysts (photocatalysts) can be used to drive water splitting – the reaction in which water is broken down into hydrogen and oxygen gases. Takata *et al.*[1] prepared photocatalysts made of highly crystalline strontium titanate that contains a small number of aluminium atoms. When the particles are suspended in water and irradiated with ultraviolet light, electrons and holes (positively charged quasiparticles) are produced, and travel to different facets of the crystalline particle. The authors selectively deposited appropriate co-catalysts on the facets to promote hydrogen production (using the electrons) at the electron-collecting facets, and oxygen production (using the holes) at the hole-collecting facets. The selective migration of electrons and holes to different reaction sites contributed to the almost perfectly efficient conversion of light to hydrogen molecules by the photocatalyst.

"This combination of complex mitigation strategies proved highly successful."

Takata *et al.* focus on strontium titanate ($SrTiO_3$) – one of the first materials found to split water photocatalytically, as reported[5] in 1977. Strontium titanate produces electron–hole pairs by absorbing ultraviolet light. Because the Sun's intensity is highest in the visible-light range, it is unlikely that UV-driven catalysts will enable sustainable hydrogen production on a large scale. However, strontium titanate is an excellent model system for studying the influence of photocatalyst parameters on quantum efficiency (both EQE and IQE), because the mechanisms that cause efficiency losses in this material are well understood, and strategies for mitigating the losses have been proposed.

The authors used a combination of approaches to address specific loss mechanisms. One such mechanism is charge recombination, a process in which electrons and holes recombine before they can take part in water splitting. Takata and colleagues suppressed recombination in several ways. The first approach was to improve the crystallinity of the photocatalyst particles, thereby reducing the number of lattice defects. Another method was to reduce the number of chemical defects in the crystal lattice using aluminium doping – a process in which small quantities of aluminium atoms are incorporated into the lattice. These two approaches work because any defect (a lattice defect or a chemical defect) can act as a potential centre for recombination[6].

Takata and colleagues also took advantage of the fact that electrons and holes in their strontium titanate crystals collect at different crystal facets – a feature that further suppresses charge recombination. The authors selectively deposited appropriate co-catalysts on the facets to promote hydrogen production at the electron-collecting facets, and oxygen production at the hole-collecting facets (Fig. 1); this approach was previously proposed[7] and developed[8] by other research groups. Finally, the authors prevented an unwanted side reaction (the oxygen-reduction reaction) by encasing the rhodium co-catalyst

© 2020 Springer Nature Limited. All rights reserved.

for the hydrogen-producing reaction in a protective shell of a chromium compound.

This combination of complex mitigation strategies proved highly successful: the authors reported EQEs of up to 96% when their photocatalysts were irradiated with light in the wavelength range of 350–360 nanometres. This is excellent news, because it means they have designed an almost perfect photocatalyst – the IQE must be between 96% and 100%.

This is a spectacular result for several reasons, even though strontium titanate is 'just' a model system for visible-light photocatalysts. First, it demonstrates that experiments can be designed in which EQEs come close to IQEs within an acceptable error margin of less than 4%. Improved experimental set-ups in which measured EQEs are very near to IQEs should facilitate the comparison of photocatalysts and therefore accelerate progress in this field.

Second, it proves that the combination of design strategies used by the authors can indeed eliminate efficiency losses associated with recombination. It is to be expected that the strategies used to improve the efficiency of strontium titanate will also apply to photocatalysts driven by visible light – and could therefore enable the conversion of solar energy to hydrogen with efficiencies of about 10%.

Finally, and most importantly, Takata and colleagues' findings will inspire and encourage other researchers to continue their work on photocatalysts. One of the authors of the work, Kazunari Domen, published his first paper[9] on the use of strontium titanate as a photocatalyst in 1980. This shows the timescale needed for success in this area. Although we do not yet have a route for the sustainable and economically viable production of hydrogen, we stand a good chance of finding one in the next few decades. This paper vouches for it.

Simone Pokrant is in the Department of Chemistry and Physics of Materials, University of Salzburg, Salzburg A-5020, Austria.
e-mail: simone.pokrant@sbg.ac.at

1. Takata, T. *et al. Nature* **581**, 411–414 (2020).
2. Abbott, D. *Proc. IEEE* **98**, 42–66 (2010).
3. Ochoa Robles, J., De-León Almaraz, S. & Azzaro-Pantel, C. in *Hydrogen Supply Chains* 8–11 (Academic, 2018).
4. Pinaud, B. A. *et al. Energy Environ. Sci.* **6**, 1983–2002 (2013).
5. Wrighton, M. S., Wolczanski, P. T. & Ellis, A. B. *J. Solid State Chem.* **22**, 17–29 (1977).
6. Ham, Y. *et al. J. Mater. Chem. A* **4**, 3027–3033 (2016).
7. Giocondi, J. L. & Rohrer, G. S. *J. Am. Ceram. Soc.* **86**, 1182–1189 (2003).
8. Mu, L. *et al. Energy Environ. Sci.* **9**, 2463–2469 (2016).
9. Domen, K., Naito, S., Soma, M., Onishi, T. & Tamaru, K. *J. Chem. Soc. Chem. Commun.* 543–544 (1980).

Molecular biology

Evolution of a molecular machine

Michael Berenbrink

The multi-subunit protein haemoglobin relies on complex interactions between its components to function properly. Analysis of ancient precursors suggests that its evolution from a simple monomer involved only a few steps. **See p.480**

The oxygen-transporting protein haemoglobin has undergone repeated adaptations as animals evolved to conquer new environments – from the depths of the oceans[1] to high mountain ranges[2]. These adaptations relied on changes in the long-range interactions between oxygen-binding sites buried in the protein's subunits, and between these regions and binding sites for a multitude of small effector molecules on the protein's surface[3]. How did this complex molecular machine, which can respond so exquisitely to available levels of both oxygen and several other effector molecules, come into being? On page 480, Pillai *et al.*[4] reconstruct the stepwise evolution of haemoglobin from precursors that existed more than 400 million years ago.

Almost nothing was previously known about how the four-subunit (tetrameric) form of haemoglobin that is found in modern-day jawed vertebrates evolved from ancient monomers. Tetrameric haemoglobin consists of two α- and two β-subunits. Pillai *et al.* computationally reconstructed an evolutionary tree to chart the protein's ancient history, using the amino-acid sequences of a large collection of the closely related vertebrate globin proteins, which exist as either monomers or tetramers.

"Pillai and colleagues' work serves as one of the clearest examples so far of how such complexity can arise."

The authors' tree was constructed taking into account that amino-acid substitutions a given protein shares with close relatives tend to have originated in more-recent common ancestors than have those it shares with more-distant relatives. The reconstructed evolutionary tree indicates that multiple rounds of gene duplication and subsequent divergence gave rise to the globin family and, by way of several ancestral proteins, to tetrameric haemoglobin (Fig. 1).

What is special about the study is that Pillai and colleagues went on to resurrect several of these extinct ancestral proteins, generating them from the amino-acid sequences predicted by the tree. The group then tested these proteins' functions.

First, Pillai and colleagues analysed whether each ancestral protein could form dimers and tetramers of like or unlike subunits. The earliest protein – a common ancestor of haemoglobin and the monomeric globin protein myoglobin, named AncMH by the authors – exists only as a monomer. A later protein, named Ancα/β, which is the ancestor of all existing haemoglobin subunits, forms homodimers when expressed at high levels. The authors' tree indicates that Ancα/β underwent gene duplication to produce two proteins: the ancestors of all existing α- or β-subunits, which the group respectively named Ancα and Ancβ. These proteins also form homodimers, or even homotetramers, when expressed alone. However, when the two are expressed together in equal proportions, they can form heterodimers, which then further align to yield haemoglobin tetramers.

The group next investigated the oxygen-binding affinity of the ancestral proteins, along with their oxygen cooperativity (the ability of oxygen-binding subunits to interact with one another) and their 'allosteric' regulation by a potent, artificial effector molecule, inositol hexaphosphate (IHP). They found that only Ancα and Ancβ – when expressed together at high concentrations – show similar oxygen-binding affinity, cooperativity and allosteric regulation to today's haemoglobin protein. These features are shared by all living jawed vertebrates, but are absent or achieved in a different way in jawless vertebrates, whose haemoglobin proteins are of more ancient origin. This indicates that the basic functions of jawed-vertebrate haemoglobin had already evolved in a common ancestor of these animals but at some time after the split with jawless vertebrates.

Next, Pillai *et al.* modelled the stepwise changes in α- and β-subunit interfaces that might have allowed Ancα and Ancβ first to form heterodimers with one another, and later heterotetramers from pairs of such dimers. The modelling indicated that strikingly few amino-acid substitutions might have been needed to transform a simple monomeric

© 2020 Springer Nature Limited. All rights reserved.

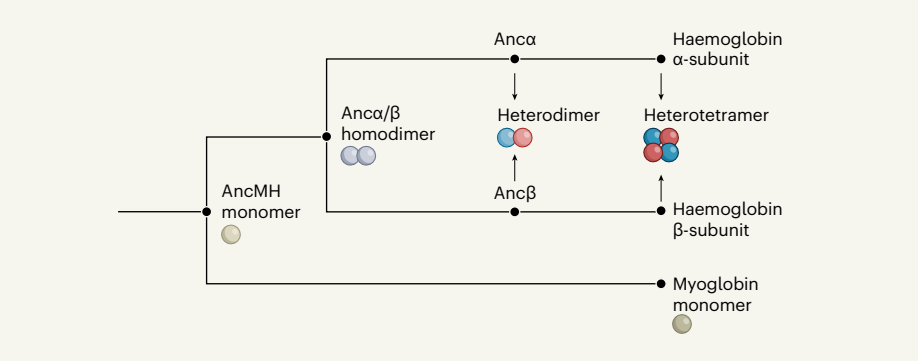


Figure 1 | Key steps in the evolution of the tetrameric haemoglobin protein. In jawed vertebrates, haemoglobin exists as a tetramer, formed from two α- and two β-subunits. Pillai *et al.*[4] resurrected extinct ancestors of haemoglobin using predicted amino-acid sequences to reconstruct the protein's evolution. The authors showed that the protein AncMH, the last common ancestor of haemoglobin and the related protein myoglobin, existed as a monomer. Duplication of the gene that encoded AncMH, and subsequent divergence into two genes, produced monomeric myoglobin and the ancestor of haemoglobin, Ancα/β, which forms a homodimer. Further gene duplication of Ancα/β and subsequent divergence yielded the ancestors of the α- and β-subunits, dubbed Ancα and Ancβ. These two subunits evolved an interface that allowed the formation of heterodimers. A few further changes in amino-acid residues generated a second interface that allowed the assembly of modern-day α- and β-subunits into haemoglobin heterotetramers.

oxygen-binding protein similar to myoglobin (whose oxygen binding is non-cooperative and almost totally unaffected by allosteric effector molecules[5]) into haemoglobin. Importantly, the researchers verified the results of their model by generating proteins that carried mutations of the amino-acid residues identified, and showing that heterotetramer formation was disrupted.

The authors' work shows how natural selection, acting on pre-existing biophysical protein properties, can, in just a few evolutionary steps, create multimeric structures that have complex functions. Most cellular processes involve the action of protein multimers, and Pillai and colleagues' work serves as one of the clearest examples so far of how such complexity can arise during protein evolution.

There are inevitable uncertainties in these kinds of reconstruction of the deep past of life, because the accuracy of such reconstructions relies on the proteins under consideration having several specific properties[6]. The proteins should, ideally, show small overall rates of amino-acid sequence divergence from one another, have thoroughly known and well-supported evolutionary relationships to each other, and exhibit a dense evolutionary branching pattern over the time period(s) of interest. Finally, a detailed knowledge of structure–function relationships is essential. It would be difficult to reconstruct with any confidence ancestral sequences and functions for proteins that do not fulfil all or any of these conditions. However, haemoglobin is well suited for this type of study for several reasons. For instance, a wealth of comparative data on globin function across vertebrates has accumulated over the past 100 years[7]. We have intimate knowledge of haemoglobin structure–function relationships[3,8,9]. In addition, there is an ever-expanding pool of globin sequence information, thanks to genome-sequencing projects in diverse organisms (these efforts will also benefit similar studies on other proteins).

Pillai and colleagues' study is sure to raise several follow-up questions and to spark further research. For instance, the authors used the artificial effector IHP in their experiments, but the binding sites for physiologically relevant effectors of haemoglobin oxygen affinity, such as hydrogen ions, only partly overlap with — and in some cases are quite different from — the IHP binding site[3,8,9]. Some evidence[1] suggests that mechanisms by which hydrogen ions modify haemoglobin oxygen affinity have evolved independently multiple times in vertebrates. This would make the picture much more complicated than can be assessed using IHP.

It will be interesting to probe the evolutionary origins of the regulation of haemoglobin oxygen binding by other effectors, such as carbon dioxide or physiologically relevant organic phosphates including ATP and 2,3-bisphosphoglycerate. In doing so, we could examine, for instance, how haemoglobin regulation changed as demands on the body's oxygen-transport system rose during the evolution of warm-blooded, active vertebrates, or how it was affected by changes over geological time in atmospheric oxygen levels, which have been proposed by some to have shaped vertebrate evolution[10].

The assembly of multimeric proteins depends on specific concentrations and thus expression levels of the protein's subunits. Natural selection presumably prevents imbalanced subunit production, both to limit the costly energy expenditure involved in protein synthesis and to prevent accumulation of potentially harmful spare globin subunits, as occurs in some hereditary human blood disorders[11]. But, at some point, an increase and balancing of expression levels of haemoglobin's subunits would have been needed to enable tetramer formation. When did this occur? It has been shown[12] that the net surface charge of myoglobin acts as a molecular 'signature' that can be used to assess the expression levels of ancestral myoglobin. However, such markers are largely unknown from other globin proteins, and we do not know the ancestral expression levels of any of the reconstructed haemoglobin precursors in Pillai and colleagues' study.

Finally, as previously noted[13], one of the most fascinating frontiers in this research field might be uncovering the evolutionary history of gene regulation. This remains an open question in the evolution of the genes that encode haemoglobin's subunits.

Michael Berenbrink is in the Department of Evolution, Ecology and Behaviour, University of Liverpool, Liverpool L69 7ZB, UK.
e-mail: michael.berenbrink@liverpool.ac.uk

This article was published online on 20 May 2020.

1. Berenbrink, M., Koldkjær, P., Kepp, O. & Cossins, A. R. *Science* **307**, 1752–1757 (2005).
2. Natarajan, C. *et al. Science* **354**, 336–339 (2016).
3. Perutz, M. F. *Mol. Biol. Evol.* **1**, 1–28 (1983).
4. Pillai, A. S. *et al. Nature* **581**, 480–485 (2020).
5. Antonini, E. & Brunori, M. *Hemoglobin and Myoglobin in Their Reactions with Ligands* 21 (North-Holland, 1971).
6. Benner, S. A. in *Ancestral Sequence Reconstruction* (ed. Liberles, D. A.) Ch. 1, 3–19 (Oxford Univ. Press, 2007).
7. Krogh, A. & Leitch, I. *J. Physiol. (Lond.)* **52**, 288–300 (1919).
8. Storz, J. F. *Hemoglobin: Insights into Protein Structure, Function, and Evolution* (Oxford Univ. Press, 2019).
9. Imai, K. *Allosteric Effects in Haemoglobin* (Cambridge Univ. Press, 1982).
10. Schachat, S. R. *et al. Proc. R. Soc. B* **285**, 20172631 (2018).
11. Rund, D. & Rachmilewitz, E. *N. Engl. J. Med.* **353**, 1135–1146 (2005).
12. Mirceta, S. *et al. Science* **340**, 1234192 (2013).
13. Zuckerkandl, E. in *Ancestral Sequence Reconstruction* (ed. Liberles, D. A.) x–xi (Oxford Univ. Press, 2007).

© 2020 Springer Nature Limited. All rights reserved.

Article

A census of baryons in the Universe from localized fast radio bursts

https://doi.org/10.1038/s41586-020-2300-2

Received: 27 September 2019

Accepted: 13 March 2020

Published online: 27 May 2020

Check for updates

J.-P. Macquart[1]✉, J. X. Prochaska[2,3]✉, M. McQuinn[4], K. W. Bannister[5], S. Bhandari[5], C. K. Day[5,6], A. T. Deller[6], R. D. Ekers[1,5], C. W. James[1], L. Marnoch[5,7], S. Osłowski[6], C. Phillips[5], S. D. Ryder[7], D. R. Scott[1], R. M. Shannon[6] & N. Tejos[8]

More than three-quarters of the baryonic content of the Universe resides in a highly diffuse state that is difficult to detect, with only a small fraction directly observed in galaxies and galaxy clusters[1,2]. Censuses of the nearby Universe have used absorption line spectroscopy[3,4] to observe the 'invisible' baryons, but these measurements rely on large and uncertain corrections and are insensitive to most of the Universe's volume and probably most of its mass. In particular, quasar spectroscopy is sensitive either to the very small amounts of hydrogen that exist in the atomic state, or to highly ionized and enriched gas[4–6] in denser regions near galaxies[7]. Other techniques to observe these invisible baryons also have limitations; Sunyaev–Zel'dovich analyses[8,9] can provide evidence from gas within filamentary structures, and studies of X-ray emission are most sensitive to gas near galaxy clusters[9,10]. Here we report a measurement of the baryon content of the Universe using the dispersion of a sample of localized fast radio bursts; this technique determines the electron column density along each line of sight and accounts for every ionized baryon[11–13]. We augment the sample of reported arcsecond-localized[14–18] fast radio bursts with four new localizations in host galaxies that have measured redshifts of 0.291, 0.118, 0.378 and 0.522. This completes a sample sufficiently large to account for dispersion variations along the lines of sight and in the host-galaxy environments[11], and we derive a cosmic baryon density of $\Omega_b = 0.051^{+0.021}_{-0.025} h_{70}^{-1}$ (95 per cent confidence; $h_{70} = H_0/(70\ \mathrm{km\ s^{-1}\ Mpc^{-1}})$ and H_0 is Hubble's constant). This independent measurement is consistent with values derived from the cosmic microwave background and from Big Bang nucleosynthesis[19,20].

The Commensal Real-time ASKAP Fast Transients (CRAFT) survey on the Australian Square Kilometre Array Pathfinder (ASKAP) has commissioned a mode capable of localizing fast radio bursts (FRBs) with subarcsecond accuracy, thus enabling identification of their host galaxies and measurement of their redshifts z. ASKAP consists of 36 antennas equipped with phased array feeds, able to view 30 degrees^2 on the sky. Bursts are detected by incoherently summing the total power signal of individual beams from each of the antennas. Bursts detected in the incoherent pipeline are subsequently localized interferometrically by triggering a download of voltage data from a 3.1-s-duration ring buffer that is correlated and imaged at high time resolution to provide the localizations[15,16]. The 6-km baselines of the array yield statistical position errors of approximately $10''(S/N)^{-1}$, where the final coherent signal-to-noise ratio of the burst, S/N, exceeds 50 for any burst whose signal-to-noise ratio in the incoherent pipeline is greater than 9. The resulting statistical (thermal) uncertainties are smaller than 0.2″. Systematic errors in these positions are typically smaller than 0.5″. At $z = 0.5$, 1″ corresponds to 5 kpc which is approximately the precision needed to associate an FRB with its host galaxy while reducing the chance coincidence probability to <1% (ref. [16]).

We report the detection of four localized ASKAP bursts. Table 1 lists the burst properties, sky positions and host galaxy offsets, while Fig. 1 shows the host galaxy identifications (see also ref. [21] and Methods). Their dispersion measures (DMs) are well in excess of the 30–100 pc cm^{-3} contributions expected from the disk and halo of the Milky Way at high Galactic latitudes[13,22], with the large excesses attributable to the intergalactic medium (IGM) and gas within each burst host galaxy. Two other ASKAP-detected bursts and their host galaxies were reported previously[15,16] in addition to three other host-galaxy identifications[14,17,18].

The precise localization of a set of FRBs to their host galaxies provides the first ensemble of $\mathrm{DM_{FRB}}$ and z_{FRB} measurements. The $\mathrm{DM_{FRB}}$ measurement represents the electron density weighted by $(1+z)^{-1}$ integrating over all physical distance increments ds to a given FRB: $\mathrm{DM_{FRB}} = \int n_e\, \mathrm{d}s/(1+z)$. Physically, we expect $\mathrm{DM_{FRB}}$ to separate into four primary components:

[1]International Centre for Radio Astronomy Research, Curtin Institute of Radio Astronomy, Curtin University, Perth, Western Australia, Australia. [2]University of California Observatories–Lick Observatory, University of California, Santa Cruz, CA, USA. [3]Kavli Institute for the Physics and Mathematics of the Universe, Kashiwa, Japan. [4]Astronomy Department, University of Washington, Seattle, WA, USA. [5]Commonwealth Science and Industrial Research Organisation, Australia Telescope National Facility, Epping, New South Wales, Australia. [6]Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria, Australia. [7]Department of Physics and Astronomy, Macquarie University, North Ryde, New South Wales, Australia. [8]Instituto de Física, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile. ✉e-mail: J.Macquart@curtin.edu.au; xavier@ucolick.org

Table 1 | Properties of FRBs interferometrically localized with ASKAP

FRB	Time (UTC)[a]	DM (pc cm^{-3})	Rotation measure (rad m^{-2})	Fluence (Jy ms)[b]	Right ascension (h:min:s)[c]	Declination (degree:arcmin:arcsec)[c]	Host redshift	Offset from host nucleus (kpc)
180924	16:23:12.6265	361.42(6)	14(1)	16(1)	21:44:25.255 ± 0.006 ± 0.008	−40:54:00.10 ± 0.07 ± 0.09	0.3214	3.5 ± 0.9
181112	17:31:15.48365	589.27(3)	10.9(9)	26(3)	21:49:23.63 ± 0.05 ± 0.24	−52:58:15.4 ± 0.3 ± 1.4	0.4755	$3.1^{+15.7}_{-3.1}$
190102	05:38:43.49184	363.6(3)	110	14(1)	21:29:39.76 ± 0.06 ± 0.16	−79:28:32.5 ± 0.2 ± 0.5	0.291	$1.5^{+3.4}_{-1.5}$
190608	22:48:12.88391	338.7(5)	–	26(4)	22:16:04.75 ± 0.02 ± 0.02	−07:53:53.6 ± 0.3 ± 0.3	0.1178	7.0 ± 1.3
190611	5:45:43.29937	321.4(2)	–	10(2)	21:22:58.91 ± 0.11 ± 0.23	−79:23:51.3 ± 0.3 ± 0.6	0.378	17.2 ± 4.9
190711	01:53:41.09338	593.1(4)	–	34(3)	21:57:40.68 ± 0.051 ± 0.15	−80:21:28.8 ± 0.08 ± 0.3	0.522	$1.5^{+3.6}_{-1.5}$

The FRB detection pipeline makes use of the DBSCAN algorithm[25], as implemented by ref. [26], to mitigate RFI and reduce the frequency of false-positive FRB triggers.

[a]Burst arrival time referenced to a frequency of 1,152 MHz.

[b]Quoted errors on the last significant digit of the fluence represent a 90% confidence limit.

[c]Errors listed after the burst position represent the statistical and systematic uncertainties respectively, and are combined in quadrature for a final absolute positional uncertainty.

$$\mathrm{DM_{FRB}}(z) = \mathrm{DM_{MW,ISM}} + \mathrm{DM_{MW,halo}} + \mathrm{DM_{cosmic}}(z) + \mathrm{DM_{host}}(z) \quad (1)$$

with $\mathrm{DM_{MW,ISM}}$ the contribution from our Galactic ISM, $\mathrm{DM_{MW,halo}}$ the contribution from our Galactic halo[13], $\mathrm{DM_{host}}$ the contribution from the host galaxy including its halo and any gas local to the event, and $\mathrm{DM_{cosmic}}$ the contribution from all other extragalactic gas. Only $\mathrm{DM_{cosmic}}$, determined by its path length through the IGM and the increase in baryon density with look-back time, is expected to have a strong redshift dependence, although $\mathrm{DM_{host}}$ is weighted by $(1+z_{\mathrm{FRB}})^{-1}$ and may correlate with age, for example, if host galaxies have systematically lower mass at earlier times.

Adopting our cosmological paradigm of a flat universe with matter and dark energy, the average value of $\mathrm{DM_{cosmic}}$ to redshift z_{FRB} is:

$$\langle \mathrm{DM_{cosmic}} \rangle = \int_0^{z_{\mathrm{FRB}}} \frac{c\bar{n}_{\mathrm{e}}(z)\mathrm{d}z}{H_0(1+z)^2\sqrt{\Omega_{\mathrm{m}}(1+z)^3+\Omega_\Lambda}} \quad (2)$$

with mean density $\bar{n}_{\mathrm{e}} = f_{\mathrm{d}}\rho_{\mathrm{b}}(z)m_{\mathrm{p}}^{-1}(1-Y_{\mathrm{He}}/2)$, where m_{p} is the proton mass, $Y_{\mathrm{He}} = 0.25$ is the mass fraction of helium, assumed doubly ionized in this gas, $f_{\mathrm{d}}(z)$ is the fraction of cosmic baryons in diffuse ionized gas (this accounts for dense baryonic phases, for example, stars and neutral gas; see Methods), $\rho_{\mathrm{b}}(z) = \Omega_{\mathrm{b}}\rho_{\mathrm{c},0}(1+z)^3$, and Ω_{m} and Ω_Λ are the matter and dark energy densities today in units of $\rho_{\mathrm{c},0} = 3H_0^2/8\pi G$ where we parameterize Hubble's constant H_0 in terms of the dimensionless $h_{70} = H_0/(70\ \mathrm{km\ s^{-1}\ Mpc^{-1}})$.

The $\mathrm{DM_{MW,ISM}}$ term (equation (1)) arises primarily from the so-called warm ionized medium of the Galaxy and is estimated from a model of this ISM component[22,23]. At the high Galactic latitudes ($|b| > 33°$) of the ASKAP sample, the value is $\mathrm{DM_{MW,ISM}} \approx 30$ pc cm^{-3}. The $\mathrm{DM_{MW,halo}}$ term is not well constrained[13], but is expected to be in the range of approximately 50–100 pc cm^{-3}. Hereafter we assume $\mathrm{DM_{MW,halo}} = 50$ pc cm^{-3} and emphasize that the sum of its scatter and uncertainty are less than those of $\mathrm{DM_{cosmic}}$ and $\mathrm{DM_{host}}$, which we discuss below.

Figure 2 shows the theoretical curve for $\langle \mathrm{DM_{cosmic}} \rangle$ versus z_{FRB} for the Planck15 cosmology[20] and a model estimate of the scatter (90% interval) due to statistical variations in foreground cosmic structure (see Methods). Overplotted on the model are the estimated $\mathrm{DM_{cosmic}}$ and measured z_{FRB} values for all arcsecond-localized FRBs. We have estimated $\mathrm{DM_{cosmic}}$ by subtracting the following from the measured $\mathrm{DM_{FRB}}$ value: $\mathrm{DM_{MW,ISM}}$ from the Galactic ISM model; our assumed $\mathrm{DM_{MW,halo}}$ contribution; and an ansatz of $\mathrm{DM_{host}} = 50/(1+z)$ pc cm^{-3} estimated from theoretical work and informed from the analysis below. We ignore FRB 121102 and FRB 190523 in most of the analysis that follows because of selection bias in their discovery, FRB 180916 owing to its low Galactic latitude (see Methods), and FRB 190611 because of its tentative association with a host galaxy (see Methods). The five ASKAP FRBs that remain comprise what we term the gold-standard sample. The agreement between model and data is striking. Effectively, the FRB measurements confirm the presence of baryons with the density estimated from the cosmic microwave background (CMB) and Big Bang nucleosynthesis (BBN), and these five measurements are consistent with all the missing baryons being present in the ionized IGM.

This result motivated us to quantitatively test for the consistency of $\Omega_{\mathrm{b}}h_{70}$ with CMB and BBN measurements, simultaneously determining the uncertain host-galaxy contributions to $\mathrm{DM_{FRB}}$ as well as the sightline-to-sightline variance in dispersion owing to the IGM. We do this by analysing the joint likelihood of our sample of five (seven, including FRBs 190523 and 190611) $\mathrm{DM_{FRB}}$, z_{FRB} measurements against a four parameter model: one parameter for the large-scale structure scatter in $\mathrm{DM_{cosmic}}$, two parameters for $\mathrm{DM_{host}}$ (a mean and a scatter), and $\Omega_{\mathrm{b}}h_{70}$. Our model for the contributions to $\mathrm{DM_{FRB}}$ starts with equations (1) and (2), and we develop parametric models for $\mathrm{DM_{host}}$ and the intrinsic scatter in $\mathrm{DM_{cosmic}}$. We again fix $\mathrm{DM_{MW,halo}} = 50$ pc cm^{-3} and adopt the Galactic ISM model[22] for $\mathrm{DM_{MW,ISM}}$. Uncertainty in these values can be absorbed into our model for $\mathrm{DM_{host}}$.

For $\mathrm{DM_{cosmic}}$, our model accounts for scatter in the electron column from foreground structures, which is largely caused by random variation in the number of haloes a given sightline intersects. Cosmological simulations show that this variation is sensitive to the extent to which galactic feedback redistributes baryons around galactic haloes[11,16] and that the fractional standard deviation of the cosmic DM equals approximately[11] $Fz^{-1/2}$ for $z < 1$, where the parameter F quantifies the strength of the baryon feedback (0.1 being strong feedback and 0.4 being weak). Stronger feedback corresponds to situations in which feedback processes expel baryons to larger radii from their host galaxies or where more massive haloes are evacuated by such feedback. The formalism incorporates the effect of large-scale structure associated with voids and the intersection of sightlines with clusters. We find that a one-parameter model based on a physically motivated shape for the probability distribution of $\mathrm{DM_{cosmic}}$ given F provides a successful description of a wide range of cosmological simulations (see Methods). Our form for the distribution is strongly asymmetric towards lower redshifts, admitting large $\mathrm{DM_{cosmic}}$ values that we find are important in the estimation of $\Omega_{\mathrm{b}}h_{70}$.

We chose our model for $\mathrm{DM_{host}}$ to follow a log-normal distribution, characterized by a median $\exp(\mu)$ and logarithmic width parameter σ_{host} such that the standard deviation of the distribution is $\exp(\mu)e^{\sigma_{\mathrm{host}}^2/2}(e^{\sigma_{\mathrm{host}}^2}-1)^{1/2}$. We do not attempt to incorporate redshift-dependent evolution in the host-galaxy dispersion contribution, but we do scale the distribution of $\mathrm{DM_{host}}$ by the factor $(1+z_{\mathrm{host}})^{-1}$ applicable to a parcel of plasma at redshift z_{host} so that $\mathrm{DM_{host}}$ is interpreted as the dispersion in the rest frame of the host galaxy. Our choice of a log-normal distribution is conservative in that it allows for a tail

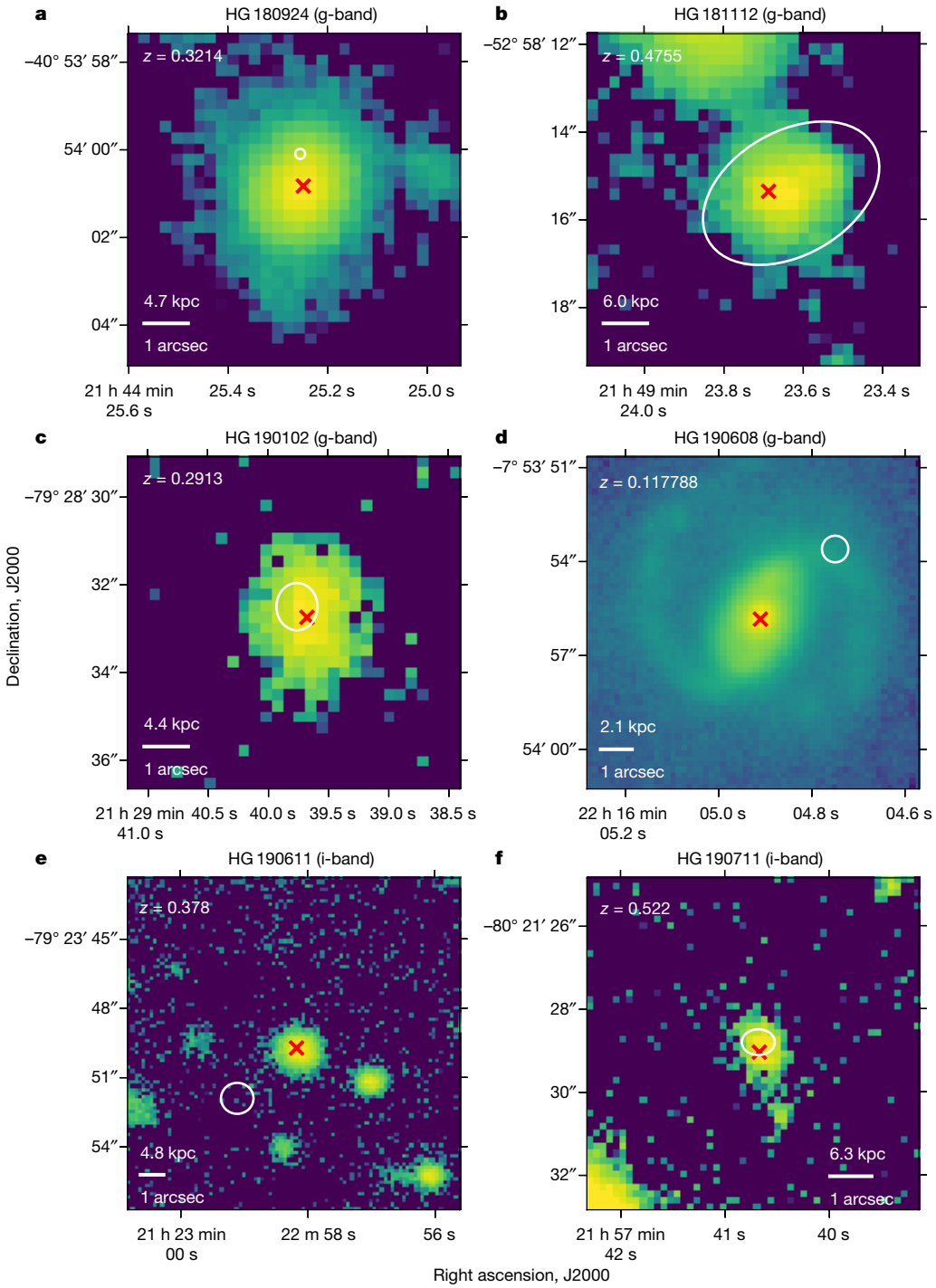


Fig. 1 | Locations of FRBs relative to their host galaxies. a–f, Optical images of the host galaxies of six FRBs localized by ASKAP, including the four new bursts reported here. **a**, FRB 180924; **b**, FRB 181112; **c**, FRB 190102; **d**, FRB 190608; **e**, FRB 190611; and **f**, FRB 190711. HG, host galaxy. White ellipses denote the 90% confidence region of each burst position, including statistical uncertainty and phase referencing errors, while the red crosses mark the measured centroids of each host galaxy. The identification of the host galaxy of FRB 190611 is tentative. **a–d**, Deep VLT g-band images; **e**, **f**, deep GMOS i-band images.

extending to large positive values, which may not be present in our sample given our selection criteria (see Methods) and the burst locations relative to the host stellar surface density. We explore $\mathrm{DM_{host}}$ distributions with median values in the range $\exp(\mu) = 20–200\ \mathrm{pc\ cm^{-3}}$ and σ_{host} in the range 0.2–2.0.

Our final analysis compares the relative likelihood of models in a four-parameter space ($\Omega_{\mathrm{b}}h_{70}$, F, σ_{host}, μ); see Methods for a Bayesian approach that yields similar constraints. Marginalizing the other parameters over ranges restricted by other physical constraints, we derive the constraints on $\Omega_{\mathrm{b}}h_{70}$ shown in Fig. 3 using our five-FRB gold-standard sample. The results are fully consistent with the joint CMB + BBN estimations and with only five (seven) burst redshifts the experiment yields a precision of $\sigma(\Omega_{\mathrm{b}}h_{70})/\Omega_{\mathrm{b}}h_{70} = 0.31$ (0.28) at the 68% confidence level, with F marginalized over the range [0.09, 0.32] (see Methods). This quantitative result for $\Omega_{\mathrm{b}}h_{70}$ substantiates our inference from the DM–z relation in Fig. 2 that the FRB ensemble has

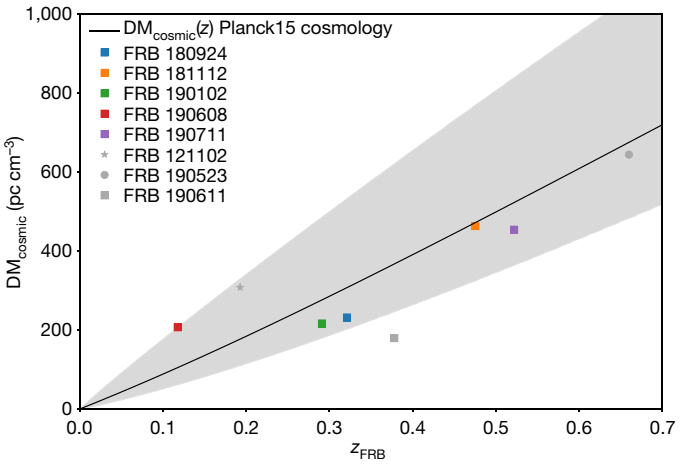


Fig. 2 | The DM–redshift relation for localized FRBs. Data points are estimations of the cosmic dispersion measure (DM_{cosmic}) versus FRB redshift (z_{FRB}) for all current arcsecond- and subarcsecond-localized FRBs. The DM_{cosmic} values are derived by correcting the observed dispersion measure DM_{FRB} for the estimated contributions from our Galaxy and the FRB host galaxy (the latter assumed here to be $50(1+z)^{-1}$ pc cm^{-3}; see text for details). Coloured points represent the gold-standard sample on which our primary analysis is based. The solid line denotes the expected relation between DM_{cosmic} and redshift for a universe based on the Plank15 cosmology (that is, $\Omega_b = 0.0486$ and $H_0 = 67.74$ km s^{-1} Mpc^{-1}). The shaded region encompasses 90% of the DM_{cosmic} values from a model for ejective feedback in Galactic haloes that is motivated by some simulations (with $F = 0.2$ in equation (4) in Methods), illustrating that the observed scatter is largely consistent with the scatter from the IGM.

resolved the missing baryons problem. The ratio of our estimated Ω_b to that from CMB and BBN measurements is $1.1^{+0.5}_{-0.6}\ h_{70}$. Formally, we exclude $\Omega_b h_{70} < 0.02$ (0.01) at the 98.6% (99.8%) confidence level. This constraint should improve considerably in the near term as ASKAP and other facilities acquire a larger sample of bursts with redshifts.

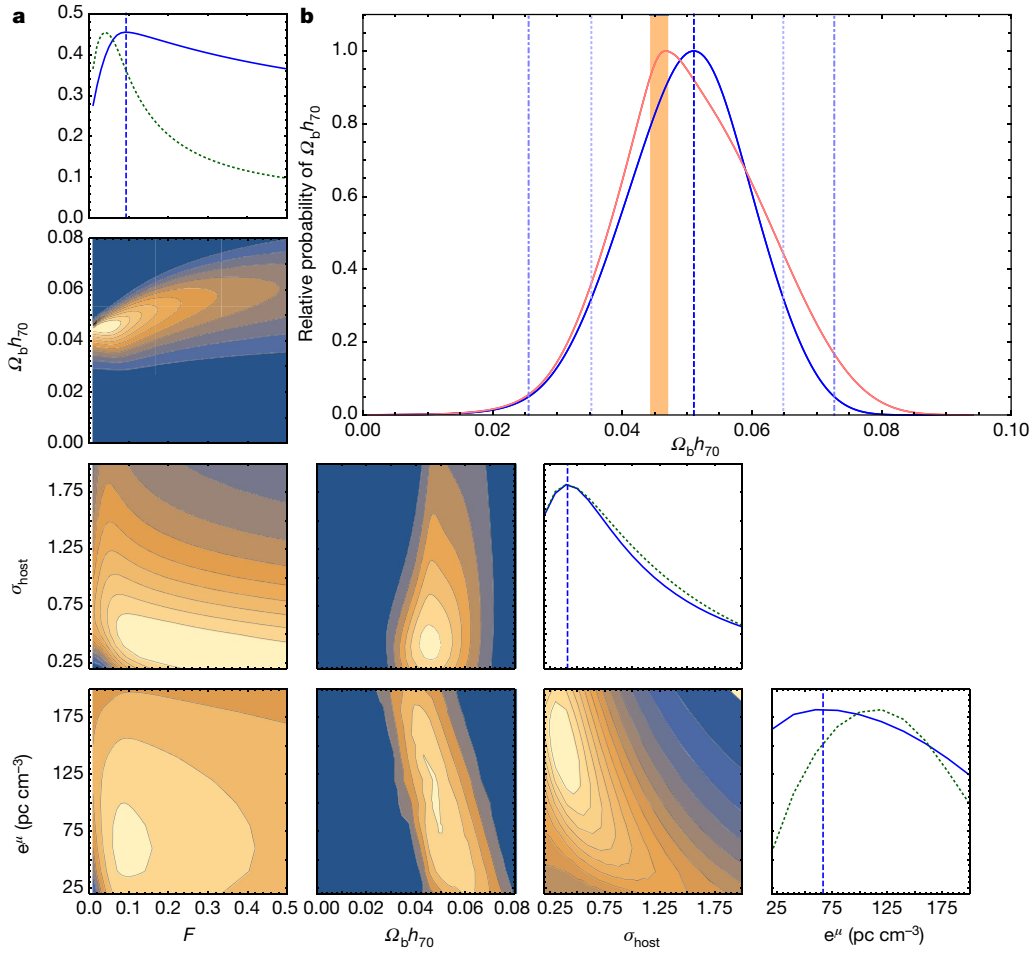


Fig. 3 | The density of cosmic baryons derived from the FRB sample. The constraints on the IGM parameters $\Omega_b h_{70}$ and F, and the host galaxy parameters $\exp(\mu)$ and σ_{host}, for a log-normal DM distribution are derived using the five gold-standard bursts (as described in the text and Methods). **a**, Corner plots displaying the probability of a given value of F, $\exp(\mu)$ or σ_{host} relative to its most likely value, and marginalized over the other parameters: heavy dashed lines represent the most likely values in each case. The green, dotted lines in the corner plots of F, e^{μ} and σ_{host} denote the relative likelihood of these parameters when $\Omega_b h_{70}$ is constrained to the value set by the CMB + BBN measurements. The contours displayed are in increments of 10% of the peak value. **b**, Magnified view of the corner plot for $\Omega_b h_{70}$ where the orange shaded region denotes the range to which $\Omega_b h_{70}$ is confined by CMB + BBN measurements. The dotted and dot-dashed lines represent the 68% and 95% confidence intervals of each parameter, respectively. The distribution of Ω_b is alternatively marginalized over the range of F indicated by cosmological simulations, [0.09, 0.32] (blue curve; see Methods), and over the entire range of F [0,0.5] investigated here (red curve).

Additionally, analysis of our gold-standard sample mildly favours a median host galaxy contribution of about 100 pc cm^{-3} with a factor of two dispersion around this value ($\sigma_{\rm host} \approx 1$). This quantifies our result that the host contributions are sufficiently small to not compromise the use of FRBs for cosmology and IGM science. Even with our current small sample, we are beginning to constrain viable models for the redistribution of the cosmic baryons by galactic feedback. If we adopt a prior on $\Omega_{\rm b}h_{70}$ from the CMB, BBN and supernovae surveys[24], we find $F = 0.04^{+0.26}_{-0.04}$ (68% confidence), and if we further include FRB 190523 and FRB 190611 we find $F = 0.23^{+0.27}_{-0.12}$ (see Methods). A factor of two smaller error would start to differentiate between viable feedback scenarios (as discussed further in Methods), suggesting that FRBs have not only revealed that all the baryons are present but—with modestly larger samples—could constrain where they lie.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2300-2.

1. Fukugita, M., Hogan, C. J. & Peebles, P. J. E. The cosmic baryon budget. *Astrophys. J.* **503**, 518–530 (1998).
2. Cen, R. & Ostriker, J. P. Where are the baryons? II. Feedback effects. *Astrophys. J.* **650**, 560–572 (2006).
3. Shull, J. M., Smith, B. D. & Danforth, C. W. The baryon census in a multiphase intergalactic medium: 30% of the baryons may still be missing. *Astrophys. J.* **759**, 23 (2012).
4. Nicastro, F. et al. Observations of the missing baryons in the warm–hot intergalactic medium. *Nature* **558**, 406–409 (2018).
5. Tripp, T. M. et al. The heavy-element enrichment of Lyα clouds in the Virgo supercluster. *Astrophys. J.* **575**, 697–711 (2002).
6. Tumlinson, J. et al. The large, oxygen-rich halos of star-forming galaxies are a major reservoir of galactic metals. *Science* **334**, 948–952 (2011).
7. Prochaska, J. X., Weiner, B., Chen, H. W., Mulchaey, J. & Cooksey, K. Probing the intergalactic medium/galaxy connection. V. On the origin of Lyα and O VI absorption at z < 0.2. *Astrophys. J.* **740**, 91 (2011).
8. Hojjati, A. et al. Cross-correlating Planck tSZ with RCSLenS weak lensing: implications for cosmology and AGN feedback. *Mon. Not. R. Astron. Soc.* **471**, 1565–1580 (2017).
9. de Graaff, A., Cai, Y.-C., Heymans, C. & Peacock, J. A. Probing the missing baryons with the Sunyaev-Zel'dovich effect from filaments. *Astron. Astrophys.* **624**, A48 (2019).
10. Eckert, D. et al. Warm–hot baryons comprise 5–10 per cent of filaments in the cosmic web. *Nature* **528**, 105–107 (2015).
11. McQuinn, M. Locating the "missing" baryons with extragalactic dispersion measure estimates. *Astrophys. J.* **780**, L33 (2014).
12. Macquart, J. P. et al. Fast transients at cosmological distances with the SKA. In *Proc. Advancing Astrophysics with the Square Kilometre Array (AASKA14)* (eds Bourke, T. L. et al.) 55 (Proceedings of Science, 2015).
13. Prochaska, J. X. & Zheng, Y. Probing Galactic haloes with fast radio bursts. *Mon. Not. R. Astron. Soc.* **485**, 648–665 (2019).
14. Chatterjee, S. et al. A direct localization of a fast radio burst and its host. *Nature* **541**, 58–61 (2017).
15. Bannister, K. W. et al. A single fast radio burst localized to a massive galaxy at cosmological distance. *Science* **365**, 565–570 (2019).
16. Prochaska, J. X. et al. The low density and magnetization of a massive galaxy halo exposed by a fast radio burst. *Science* **366**, 231–234 (2019).
17. Ravi, V. et al. A fast radio burst localized to a massive galaxy. *Nature* **572**, 352–354 (2019).
18. Marcote, B. et al. A repeating fast radio burst source localized to a nearby spiral galaxy. *Nature* **577**, 190–194 (2020).
19. Cooke, R. J., Pettini, M. & Steidel, C. C. One percent determination of the primordial deuterium abundance. *Astrophys. J.* **855**, 102 (2018).
20. Planck Collaboration. Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.* **594**, A13 (2016).
21. Bhandari, S. et al. The host galaxies and progenitors of fast radio bursts localized with the Australian Square Kilometre Array Pathfinder. *Astrophys. J.* (in the press).
22. Cordes, J. M. & Lazio, T. J. W. NE2001.I. A new model for the galactic distribution of free electrons and its fluctuations. Preprint at https://arxiv.org/abs/astro-ph/0207156 (2002).
23. Gaensler, B. M., Madsen, G. J., Chatterjee, S. & Mao, S. A. The vertical structure of warm ionised gas in the Milky Way. *Publ. Astron. Soc. Aust.* **25**, 184–200 (2008).
24. Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. Preprint at http://arXiv.org/abs/1807.06209 (2018).
25. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Second Int. Conf. on Knowledge Discovery and Data Mining (KDD-96)* 226–231 (AAAI Press, 1996).
26. Novikov, A. PyClustering: data mining library. *J. Open Source Softw.* **4**, 1230 https://doi.org/10.21105/joss.01230 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Sample selection

Analogously to cosmological studies of the distance ladder with supernovae, we wish to establish a strict set of criteria for the FRB sample to minimize biases while maximizing statistical power. On the latter point, we wish to construct the largest sample while avoiding events whose DM is dominated by non-cosmological effects (that is, host or Galactic gas). Regarding bias, the greatest concern is the association of a host galaxy with a given FRB on the basis of the DM–z relation, that is, adopting this relation as a prior to establish the host identity. This is a valuable practice when one aims to resolve the underlying host-galaxy population[17,27] but would bias any cosmological study. Last, one must also be cognizant of biases related to triggering on FRB events. This practice is a complex function of the FRB fluence, its DM, and the pulse properties[28].

With these issues in mind, we propose the following set of criteria (1–4) to generate a 'gold-standard sample' of FRBs for cosmological study:

1. To make a confident host-galaxy association, we require the probability of mis-identifications to be <1% without invoking the DM–z relation since to do so would introduce a bias. For this we require the 95% localization area to encompass one and only one galaxy unless multiple galaxies have a common redshift. By 'encompass' we include light from any part of the galaxy. In practice, this will require a localization to <1″ for $z > 1$, but becoming less stringent for less distant hosts. We propose an initial set of specific criteria as follows. (i) Define 95% areas for the localization and for each galaxy in the region. Call these L and G_1, G_2, and so on. (ii) Demand one and only one galaxy overlap L. The only exception is if $z_1 = z_2$. (iii) For the overlapping L and G, require that >50% of the smaller area lies within the larger. (iv) Do this for galaxies as faint as $R = 25$ (anything fainter is generally too difficult for a spectroscopic redshift anyhow).
2. The finite temporal and spectral resolution of the FRB survey causes a decline in sensitivity with increasing DM to the point that telescope resolution causes an effective threshold at $\mathrm{DM_{cutoff}}$ at which point a burst would no longer have been detectable. A conservative approach would omit any burst with $\mathrm{DM_{cutoff}}$ sufficiently low that it excludes a large (≳30%) fraction of the total probability of $p(\mathrm{DM_{total}}|z)$, on the grounds that it presents a biased probe of $p(\mathrm{DM_{total}}|z)$. Although application of this criterion presupposes a DM–z relation and its probability density function (PDF), it does so only weakly. This point is addressed in detail below in the subsection on biases in the probability distribution.

 An event detected near to the sensitivity threshold is biased in the sense that the instrumental decline in sensitivity with increasing DM dictates that any burst detected near this threshold would not have been detectable at higher DMs. Thus, for a given redshift we are biased to finding events with DMs that are under-representative of the entire DM distribution at that redshift. Thus, only more luminous bursts, whose detection S/N is sufficiently high that $\mathrm{DM_{cutoff}}$ exceeds the plausible range of $\mathrm{DM_{total}}$ at that redshift are devoid of this bias.
3. FRB events with extreme properties (for example, high RM, large temporal broadening) will be excluded to minimize the impact of host galaxy and Galactic gas.
4. A cutoff is imposed on FRBs whose expected contribution from the disk component of the Milky Way ISM is large, to avoid large uncertainties in the subtraction of the Galactic ISM DM contribution. Models of the Galactic plasma distribution[22] typically produce errors in known pulsar distances of the order of several tens of per cent (and much higher in some cases)[29,30]. To avoid DM errors in excess of about 100 pc cm^{-3} we restrict our sample to those bursts whose predicted $\mathrm{DM_{ISM}}$ values are less than 100 pc cm^{-3}. A conservative application of this criterion restricts FRB detections to sight lines at high Galactic latitude, $|b| \gtrsim 20°$.

We acknowledge that all criteria 1–4 are subject to refinement as we learn more about FRB progenitors and their host galaxies.

Regarding criterion 3, a dominant contributor to the DM variance is the circumburst environment and the ISM of the host galaxy. Although it is not possible to make a precise estimate of this component, the burst RM, the amount of Faraday rotation exhibited by linearly polarized emission caused by its propagation through a magnetized plasma, presents a means of identifying those bursts whose radiation has probably propagated through a substantial (>100 pc cm^{-3}) amount of matter in the host galaxy. For each burst the Milky Way contribution to RM for $|b| > 10°$ is small (<250 rad m^{-2}) and measurable[31] and the IGM contribution is estimated[32] to be about 1 rad m^{-2}. Galactic haloes, similarly, have been inferred to make contributions of several tens of rad m^{-2} to the RM[33] from radio-loud quasar observations, but our first analysis with an FRB[16] yields RM < 10 rad m^{-2}. A suitable cutoff due to host-galaxy ISM contamination is suggested by assuming the host-galaxy magnetic field strength is comparable to that of our Galaxy. Measurements of Faraday rotation and dispersion from pulsars in the Milky Way (see figure 3 in ref. [34]) exhibit a mean trend $\mathrm{DM} = 1.55|\mathrm{RM}|^{0.95} \equiv f_{\mathrm{DM}}(\mathrm{RM})$, where RM and DM are measured in their usual units of rad m^{-2} and pc cm^{-3} respectively. We find that the root-mean-square (r.m.s.) deviation of the actual DM values from their values predicted on the basis of this trend using |RM| are 69% of the DM (that is, the r.m.s. errors are 69% of the mean DM value: $\langle[\mathrm{DM} - f_{\mathrm{DM}}(\mathrm{RM})]^2/\mathrm{DM}^2\rangle^{1/2} = 0.69$). We further find that there is an 85% probability that the actual DM deviates from its predicted value by less than 0.9 times the actual DM value, and a 96% probability that the predicted value differs by less than 2.0 times the DM value. We therefore suggest that a cutoff criterion $|\mathrm{RM} - \mathrm{RM_{MW}}|_{\mathrm{observed}} < 100(1+z)^2$ rad m^{-2} bounds the dispersion measure to $\mathrm{DM} \lesssim 250(1+z)$ pc cm^{-3} with 85% confidence.

A similar trend observed between the DM and the temporal smearing of Galactic pulsars caused by scattering[35,36] can also be used to place upper bounds on the host contribution. Recent updates to this relation[36] indicate that, on average, a pulse smearing time, τ, less than 33 ms (2 ms) limits DM to <200 pc cm^{-3} (300 pc cm^{-3}) at 0.327 GHz (1 GHz). However the DM–τ relation exhibits ~0.8 dex variation about the trend (as discussed in the context of FRBs in the supplementary material in ref. [15]), thus requiring $\tau < 5$ ms to ensure a reasonable (~70%) confidence that the DM contribution is less than 200 pc cm^{-3}. We caution that the use of τ as an indicator of the host-galaxy DM contribution is subject to considerable uncertainty, since neither the distances to the scattering material from the bursts, nor even the nature of the turbulence responsible for the temporal smearing observed in FRBs, is well established. The estimates presented here would be invalid, for instance, if the scattering were associated with the direct burst environment rather than the ISM of the host galaxy.

Adopting all the above criteria to the current set of FRBs with redshift estimates based on their association with galaxies (Table 1), we eliminate the following sources from cosmological analysis.

FRB 121102. We exclude the repeating FRB 121102[14] from our analysis for two reasons: (a) the rotation measure of this burst is anomalously high[37], being three orders of magnitude higher than other FRBs in this sample and indicating that this burst DM is likely contaminated by an abnormally high circumburst or host galaxy contribution, and (b) its location within 2° of the Galactic plane imparts a larger and probably less well constrained DM contribution from the Milky Way relative to the high Galactic latitude bursts detected by ASKAP.

FRB 190523. We have conducted the analysis both with and without FRB 190523. The host galaxy identification[17] from the larger, 3″ × 8″ localization region, is more uncertain than the ASKAP FRB detections and was partially based on an assumed DM–redshift relation which presents a potential source of bias in our analysis.

FRB 171020. The identification of the host galaxy associated with FRB 171020 is predicated on a search volume confined to a specific distance based on an assumed DM–redshift relation[27], and is therefore excluded from the sample. Moreover, it is difficult to ascribe a numerical value to the likelihood of a correct association in this instance.

FRB 190611. Our follow-up observations for FRB 190611 identify a galaxy at J212258.0-792350 with redshift $z = 0.378$ offset by ~2″ from the current estimate of the FRB localization. The large offset (~10 kpc at that redshift) and large systematic uncertainty in the FRB localization and the presence of a closer, faint source revealed by deep GMOS i-band imaging preclude a secure association at present. As with FRB 190523, we conduct our analysis both with and without this burst in our sample.

FRB imaging and astrometry

The procedure for characterizing the position and positional uncertainty of FRBs 190102, 190608, 190611 and 190711 followed that described in the supplementary material of refs. [15,16]. For the purposes of extracting these observables, we use only the total intensity data.

For each FRB, raw voltage data for a suitable calibrator source was captured via the CRAFT pipeline in the hours following the burst detection. For FRB 190102 and FRB 190608, the source PKS 1934-638 was used, while for FRB 190611 and FRB 190711, it was PKS 0407-658. From these calibrator data and the FRB data, visibility data sets were produced using the DiFX correlator[38]. An initial coarse search for the FRB position used the DM, pulse duration, and approximate position from the incoherently summed FRB detection data, and after detection in the interferometric data a re-correlation was performed with revised position, DM, and pulse time/duration. Radio frequency interference (RFI) was mitigated for the FRB data set by subtracting visibilities from an adjacent time range surrounding the burst itself. Additionally, for each FRB a visibility data set and image was generated using the entire 3.1 s of raw voltage data, to identify background radio continuum sources whose positions could be compared to catalogue values and verify the astrometric accuracy.

Per-station frequency-dependent complex gain calibration was derived from the calibrator data set using the ParselTongue[39]-based pipeline described in ref. [15] and transferred to the FRB data sets, before imaging in the Common Astronomy Software Applications (CASA) package. Best-fit positions and uncertainties were the extracted for each source using the task JMFIT in the Astronomical Image Processing System (AIPS)[40].

Statistical uncertainties on the FRB positions were less than 0.5″ in all cases. However, as discussed in refs. [15,16], the phase referenced FRB images will be subject to a systematic positional shift resulting from the spatial and temporal extrapolation of calibration solutions. The magnitude of this systematic shift can be estimated by comparing the positions of continuum sources in the field surrounding the FRBs to their catalogue values. The accuracy to which this can be performed depends on the number of continuum sources visible in the ASKAP continuum image and their brightness, as well as the degree to which their intrinsic source structure can be modelled (or neglected). For any given continuum source, the presence of unmodelled structure will act to shift the position of the source centroid and results in a measured offset between the ASKAP and reference positions, which perturbs the actual systematic positional shift. However, the direction of such a shift depends on the source structure, and hence should not be correlated between different continuum sources. For FRB 190102 and FRB 190611, observations made with the Australia Telescope Compact Array at a comparable frequency and angular resolution to the ASKAP image minimize the impact of source structure, but for FRB 190608, we made use of the Faint Images of the Radio Sky at Twenty centimetres (FIRST) survey[41], which has angular resolution roughly twice that of the ASKAP images, and for FRB 190711 we used archival 5 GHz ATCA data.

Assuming the phase referencing errors result in a simple translation of the FRB field image, we estimate the magnitude of this offset and its uncertainty with a weighted mean of the measured offsets for each of the continuum sources in the FRB field, after discarding any sources that were resolved in either the ASKAP image or the reference image. The magnitude of the offset ranged between 0 and 1.7 arcsec, with uncertainties ranging from 0.3 to 0.6 arcsec.

Host identification and spectroscopy

The optical spectroscopy and redshift determinations for FRB 180924 and FRB 181112 have been outlined previously[15,16]. Spectroscopy of the host galaxies of FRB 190102 and FRB 190611 was conducted using the FOcal Reducer and low dispersion Spectrograph 2[42] (FORS2) on the European Southern Observatory's Very Large Telescope (VLT) on Cerro Paranal. FORS2 was configured with the GRIS_300I grism, an OG590 blocking filter, and a 1.3″ wide slit, yielding a resolution $R_{\rm FWHM} \approx 550$. For FRB 190102 2 × 600 s exposures were obtained on 2019 March 25 UT, while for FRB 190611 2 × 1,350 s exposures were taken on 2019 July 12 UT. These and associated calibration images were processed with the PypeIt software package[43] to derive flux and wavelength calibrated spectra.

For the host galaxy of FRB 190608, the optical spectrum from the seventh data release (DR7) of the Sloan Digital Sky Survey[44] (SDSS) was retrieved from the IGMSPEC database[45].

Imaging of the host galaxies of FRB 180924, FRB 181112 and FRB 190102 was undertaken using FORS2 on the VLT, while the FRB 190608 and FRB 190711 hosts were imaged with VLT/X-shooter[46]. Imaging of the host galaxies of FRB 190611 and FRB 190711 was undertaken using GMOS on Gemini-South[47], from sets of 44 and 12 images of 100 s each in the i-band, respectively.

The FORS2 images were first reduced with ESO Reflex[48], further processed in Python, and then co-added using a median combine in Montage[49]. The WCS solutions were updated with Astrometry.net[50], with further adjustments performed by comparison with Gaia[51] or Dark Energy Survey[52] positions. The X-shooter images were reduced using a custom Python pipeline making use of the package CCDPROC[53], including measures to cope with prominent fringe patterns in I-band; the images were then co-added and the astrometry adjusted with the same method as above. The GMOS images were reduced and co-added with PYRAF using standard procedures; the astrometry was adjusted with the same method stated above. Projected distances were estimated using the Javascript Cosmology Calculator[54].

Two of the FRBs in the gold-standard sample, FRB 190608 and FRB 190711, have offsets larger than 1 arcsec from the galaxy light centroid. FRB 190608, however, is a $z = 0.11$ galaxy (that is, nearby) and the chance projection is even less than 0.3%. Regarding FRB 190711 we estimate a 6.1×10^{-3} probability that an unrelated galaxy is within a region out to a distance between the galaxy centroid and the outermost edge of the FRB error circle (for the measured $R(\mathrm{AB}) = 23.7 \pm 0.2$ mag as calibrated against the SkyMapper survey), and we estimate a probability $p = 1.9 \times 10^{-3}$ for an unrelated galaxy to be within the FRB error circle but below the detection limit of $r = 25.5$ mag. The remainder of the host-galaxy associations for each FRB have a probability $P < 10^{-3}$ of a chance occurrence[16].

The radio burst dynamic spectra and host-galaxy optical spectra are shown in Extended Data Fig. 1.

Estimating $\langle \mathrm{DM_{cosmic}} \rangle$

Central to the analysis is an estimate of the average $\mathrm{DM_{cosmic}}$ value as a function of redshift and for a given cosmology, as defined in equation (2). Previous formulations[55,56] have adopted similar definitions but with less precise considerations for $f_{\rm d}(z)$, the fraction of cosmic baryons in diffuse ionized gas. Our formulation considers the redshift evolution of three dense baryonic components that will not contribute to $n_{\rm e}$: (1) stars; (2) stellar remnants (for example, white dwarfs, neutron stars); and (3) the neutral ISM of galaxies. For (1), we interpolate

the empirically estimated stellar mass density estimates[57]. For (2), we adopt the estimation of ref. [58] which is 30% of the stellar mass. For (3), we assume the mass ratio of the ISM to stars is constant from $z = 0$–1 and adopt the present-day estimate[58] of $M_{ISM}/M_* = 0.38$. The model also allows for the partial ionization of helium but this is not relevant for the FRBs considered here. All of these calculations are encoded in Python in the public FRB repository (https://github.com/FRBs/FRB). Censuses of the gas and star evolution of baryons in $z < 1$ systems constrain the error in the fraction of neutral and non-diffuse baryons (that is, $1-f_d(z)$) to about 30% at present. Thus, with this component constituting ~15% of the total baryon budget at $z \approx 0$, the correction to $\Omega_b h_{70}$ is uncertain at a level below 6%, well below the level of precision that investigation of the current FRB sample permits. We refer the reader to ref. [59] for a discussion of the constraints on $f_d(z)$ possible in future with a larger sample of FRBs.

Cosmological parameter estimation

The ASKAP FRB measurements and localizations afford a new opportunity to constrain our cosmological paradigm through estimations of $\mathrm{DM_{cosmic}}$ and z_{FRB}. The cosmic DM is governed primarily by the baryonic density Ω_b and the expansion rate of the Universe, H_0, and the fraction of baryons in the diffuse phase, $f_d(z)$. In the following, we will assume a flat cosmology with $\Omega_\Lambda = 0.691$ (Planck15). The expansion rate is dominated by this dark energy term for $z < 0.7$, so cosmological analysis of the ASKAP FRBs is not sensitive to the precise value of Ω_m and, therefore, to a close approximation, $\langle \mathrm{DM_{cosmic}} \rangle \propto \Omega_b H_0$. We therefore proceed to place a constraint on this product.

To construct a likelihood function $\mathcal{L}$ from our FRB measurements, we build a model for $\mathrm{DM_{cosmic}}$ and its uncertainty. The model is based primarily on the cosmological parameters, but it must also allow for a nuisance parameter which accounts for the DM of our Galactic halo and that of the host galaxy: $\mathrm{DM_{MW,halo}} + \mathrm{DM_{host}}$. For the former term, theoretical models informed by observation suggest $\mathrm{DM_{MW,halo}} \approx 50\ \mathrm{pc\ cm^{-3}}$ with a small dispersion[13,56], but we acknowledge that the mean value is poorly constrained. We expect the variance in these terms to be driven by $\mathrm{DM_{host}}$, which follows from the large range in DM values observed for the ISM of our Galaxy[22], even if one ignores whether FRBs occur in 'special' locations within a galaxy. Furthermore, the very high RM and the (probably related) large DM excess of FRB 121102 above $\mathrm{DM_{cosmic}}$ implies at least one FRB with a large $\mathrm{DM_{host}}$ value[37].

The PDF for $\mathrm{DM_{host}}$ has limited theoretical motivation. In the following, we assume a log-normal distribution which has two salient features: (1) it is positive definite; (2) it exhibits an asymmetric tail to large values. The latter property allows for high $\mathrm{DM_{host}}$ values that might arise from gas local to the FRB, for example, an H II region or circumstellar medium. Formally, we adopt a log-normal distribution:

$$p_{\mathrm{host}}(\mathrm{DM_{host}}|\mu, \sigma_{\mathrm{host}}) = \frac{1}{(2\pi)^{1/2}\mathrm{DM}\sigma_{\mathrm{host}}} \exp\left[-\frac{(\log \mathrm{DM} - \mu)^2}{2\sigma_{\mathrm{host}}^2}\right] \quad (3)$$

This distribution has a median value of e^{μ} and variance $e^{\mu+\sigma_{host}^2/2}(e^{\sigma_{host}^2}-1)^{1/2}$. We consider distributions with e^{μ} in the range 20–200 pc cm^{-3} and σ_{host} in the range 0.2–2.0. An illustrative set of these probability distribution functions for $\mathrm{DM_{host}}$ is shown in Extended Data Fig. 2. For consistency of interpretation of $\mathrm{DM_{host}}$ values from bursts at disparate redshifts, the probability distribution function is referenced to the rest frame of the host galaxy, so a correction $\mathrm{DM_{host}} \rightarrow \mathrm{DM_{host}}(1+z_{FRB})^{-1}$ is applied and the distribution normalized accordingly, however, in practice this redshift correction factor varies only over the range 0.7 to 0.9 in the gold-standard sample. The inferred dispersion in $\mathrm{DM_{host}}$ is consistent with the expected range of host DMs given the galaxy type, morphology and orientation on the sky and the distance of the FRB from the galaxy centre. However, we are unable to state more than this at present, and remark that present estimates of the $\mathrm{DM_{host}}$ contributions towards specific localized FRBs[15,18] with two quite different host galaxies are, respectively, in the range 30–81 pc cm^{-3} and <70 pc cm^{-3}. This suggests that any correction on this basis could be small for our sample. A further interesting aspect of our measurements is that it is beginning to place limits on these corrections.

Altogether, $\mathrm{DM_{FRB}^{model}} = \mathrm{DM_{cosmic}}(z) + \mathrm{DM_{host}} + \mathrm{DM_{MW,ISM}}$ with the last quantity estimated from NE2001 based on the FRB coordinates; given the high Galactic latitudes of the present ASKAP sample we adopt a value $\mathrm{DM_{MW,ISM}} = 30\ \mathrm{pc\ cm^{-3}}$ for these bursts. The mechanics of our treatment of the $\mathrm{DM_{host}}$, $\mathrm{DM_{MW,halo}}$ and $\mathrm{DM_{MW,ISM}}$ terms is described in greater detail in equation (6).

The model probability distribution for $\mathrm{DM_{cosmic}}$ is derived from theoretical treatments of the IGM and galaxy haloes[11,13] with σ_{DM} dominated by the physical variance in $\mathrm{DM_{cosmic}}$. Extended Data Fig. 3 shows that comparison against the analytic form (as used in other IGM-related contexts[60])

$$p_{\mathrm{cosmic}}(\Delta) = A\Delta^{-\beta}\exp\left[-\frac{(\Delta^{-\alpha}-C_0)^2}{2\alpha^2\sigma_{\mathrm{DM}}^2}\right],\ \Delta > 0 \quad (4)$$

provides an excellent match to the $\mathrm{DM_{cosmic}}$ distributions observed in our semi-analytic models and in a hydrodynamic simulation, where $\Delta \equiv \mathrm{DM_{cosmic}}/\langle \mathrm{DM_{cosmic}} \rangle$. The motivation for this form is that in the limit of small σ_{DM}, the distribution of DM should approach a Gaussian owing to the Gaussianity of structure on large scales (a non-negligible component of the variance of $p_{cosmic}(\Delta)$ comes from tens of megaparsec structures) and in the low-σ_{DM} limit the halo gas is more diffuse and so the PDF approaches a Gaussian owing to the intersection of the line of sight with more structures. Conversely, when the variance is large, this PDF captures the large skew that results from a few large structures that contribute to the DM of many sightlines. The sharp low-DM cutoff in the distribution reflects the fact that a large component of the IGM is highly diffuse, and displays much less variance than the halo-related component, thus imposing a strict lower limit to the DM. The parameter β is related to the inner density profile of gas in haloes. If the 3D density profile scales as $\rho \propto r^{-\alpha}$, $\beta = (\alpha+1)/(\alpha-1)$ such that an isothermal profile with $\alpha = 2$ has $\beta = 3$ and an inner slope of $\alpha = 1.5$ has $\beta = 5$. Such slopes are consistent with those found in numerical simulations of intrahalo gas[61,62]. The indices $\alpha = 3$ and $\beta = 3$ provide the best match to our models (although we find that $p_{cosmic}(\Delta)$ is weakly sensitive to order unity changes in these parameters, with $\beta = 3$ having the most flexibility for our $z = 0.11$ measurement relative to $\beta = 4$). We use the parameter σ_{DM} in $p_{cosmic}(\Delta)$ as an effective standard deviation even though formally the standard deviation with $\beta = 3$ diverges logarithmically. We find that σ_{DM} is closely tied to the true standard deviation when imposing motivated maximum cutoffs for Δ on the distribution. The mean of the distribution requires that $\langle \Delta \rangle = 1$, which fixes the remaining parameter C_0 in $p_{cosmic}(\Delta)$.

Extended Data Figure 3 shows models that use equation (4) for p_{cosmic} relative to numerical calculations at redshifts that span the considered range. The solid curves are the previously described semi-analytic models[11], which assume that haloes below the specified mass have been evacuated of gas, and the 'swinds' simulation of ref. [63]. The dashed curves show the function evaluated for the best-fit σ_{DM}, and the dot-dashed curves adopt the parameterization $\sigma_{DM} = Fz^{-0.5}$ and scale off the $z = 0.5$ best fit value for σ_{DM} yielding F of 0.09, 0.15 and 0.32 in our semi-analytic models in which haloes of 10^{14}, 10^{13} and 10^{11} solar masses ($M_\odot$) are evacuated of their gas. The agreement of the dot-dashed curves with the solid numerical model curves demonstrates that $\sigma_{DM} = Fz^{-0.5}$ approximates the evolution over the range of our measurements. This scaling is further motivated in the Euclidean limit, applicable for $z \ll 1$, where $\langle \mathrm{DM_{cosmic}} \rangle = n_e cz/H$ and $\sigma_{DM} \approx \mathrm{DM_{halo}}\sqrt{N}/\langle \mathrm{DM_{cosmic}} \rangle$, where n_e is the mean electron density and N is the number of haloes intersected, which is proportional to the path length probed or cz/H.

While our analytic parameterization describes the distribution of $\mathrm{DM_{cosmic}}$ both in semi-analytic models and numerical simulations, we use

the more flexible semi-analytic models to set the marginalization range in F that is used for some constraints on $\Omega_b h_{70}$. Here we argue that the considered semi-analytic models shown in Extended Data Fig. 3 span the likely range of possible feedback scenarios. These models approximate haloes as retaining their gas in a manner that traces the dark matter above some mass threshold. This approximates the picture in many simulations[61,62] and analytic models[64,65] in which the fraction of halo gas retained is a strongly increasing function of halo mass before saturating at unity. Furthermore, gas that is outside haloes is less effective at contributing variance: take the example where gas is distributed out to a distance R around a halo. The probability a sightline intersects this gas scales as R^2, leading to less shot noise for larger R, while the contribution of each individual system scales as R^{-2}, leading to a smaller contribution for larger R. This picture motivates the semi-analytic model's approximation that ejected gas diffusely traces large-scale structure[11].

Simulations and models generally find that haloes below threshold masses in the range of about $(10^{11}–10^{13})M_\odot$ are evacuated of gas[61,62,64,65], although some implementations of stellar quasar feedback can result in different predictions[66]. Halo gas in $M > 10^{14}M_\odot$ haloes is constrained by X-ray observations to mostly reside within such haloes[67]. Our strongest feedback model, in which $F = 0.09$, pushes up against this observational limit. Our model with the weakest feedback assumes that dwarf-galaxy-sized haloes with $10^{11}M_\odot$ retain their gas and yields $F = 0.32$ (and we find that F is just marginally larger if the $10^{10}M_\odot$ haloes of the smallest dwarf galaxies retain their gas, haloes that would be just massive enough to overcome the pressure of the IGM and retain their gas[68]). Thus our models span the range of likely feedback scenarios.

Given this semi-analytic formalism, we proceed to estimate the model likelihood by computing the joint likelihoods of all FRBs:

$$\mathcal{L} = \prod_{i=1}^{N_{\rm FRBs}} P_i(\mathrm{DM'_{FRB}}|z_i) \tag{5}$$

where $P_i(\mathrm{DM'_{FRB}}|z_i)$ is the probability of the total observed $\mathrm{DM_{FRB}}$ corrected for the Galaxy:

$$\mathrm{DM'_{FRB}} \equiv \mathrm{DM_{FRB}} - \mathrm{DM_{MW,ISM}} - \mathrm{DM_{MW,halo}} = \mathrm{DM_{host}} + \mathrm{DM_{cosmic}} \tag{6}$$

For a burst at a given z_i and the model parameters we have:

$$P_i\left(\mathrm{DM'_{FRB},}i|z_i\right) = \int_0^{\mathrm{DM'_{FRB}}} p_{\rm host}(\mathrm{DM_{host}}|\mu,\sigma_{\rm host}) p_{\rm cosmic}\left(\mathrm{DM'_{FRB},}i - \mathrm{DM_{host}}, z_i\right)\mathrm{dDM_{host}} \tag{7}$$

and $p_{\rm host}(\mathrm{DM_{host}}|\mu,\sigma_{\rm host})$ the PDF for $\mathrm{DM_{host}}$. With the likelihood function defined we construct a grid of $\Omega_b H_0$, F, μ and $\sigma_{\rm host}$ values and marginalize over the last three to obtain the constraint on $\Omega_b h_{70}$. These results are presented in Fig. 3 in the main text for the gold-standard sample. Extended Data Fig. 4 presents the results of the same analysis when FRBs 190523 and 190611 are included in the data set.

To place confidence intervals on $\Omega_b H_0$ and $\sigma_{\rm DM}$, we use the likelihood ratio test statistic $\mathcal{D}$:

$$\mathcal{D}(\Omega_b h_{70}, F, \mu, \sigma_{\rm host}) = 2\log\mathcal{L}_{\rm max} - 2\log\mathcal{L}(\Omega_b h_{70}, F, \mu, \sigma_{\rm host}) \tag{8}$$

where $\mathcal{L}_{\rm max}$ is the maximum value of $\mathcal{L}$, that is, for parameters maximizing the likelihood. According to Wilks' theorem, for a sufficiently large number of FRBs, $\mathcal{D}$ will be distributed according to a χ^2_n distribution with $n = 4$ degrees of freedom[69]. If the cumulative distribution function of the $\chi^2_4(x)$ distribution is $\mathrm{CDF}(x)$, solving $\mathrm{CDF}(x) = p$ constrains the $\Omega_b h_{70}, F, \mu, \sigma_{\rm host}$ parameter space to the region $\mathcal{D} \leq x$ at confidence level (C.L.) p.

Uncertainties in these confidence estimates are probably dominated by systematic effects in the sample selection, and small number statistics. To test both, we extend the gold-standard sample of five bursts to include FRBs 190523 and 190611. The resulting analysis is shown in Extended Data Fig. 4. Compared with Fig. 3, the inclusion of two further bursts shifts the maximum-likelihood estimate for $\Omega_b h_{70}$ at 68% C.L. from $0.051^{+0.014}_{-0.015}$ to $0.042^{+0.011}_{-0.012}$, that is, consistent with the original uncertainties. This does not mean that there is no systematic bias, nor that Wilks' theorem holds precisely for our sample, but rather that any such effects are minor compared to the inherent uncertainties from our small sample size of localized bursts.

Accounting for biases in $P(\mathrm{DM},z)$

The cosmological evolution of the FRB population, and its intrinsic luminosity function, can strongly influence the observed/expected distribution of FRBs in redshift–DM space[70], $P(\mathrm{DM},z)$. We therefore perform our likelihood maximization over $P(\mathrm{DM}|z)$ only. This discards the information contained in the redshifts of our detected FRBs, but makes the procedure more robust against factors influencing the redshift distribution.

The remaining bias comes from changing sensitivity as a function of DM. This can be either direct, through DM-smearing within frequency channels, or indirect, through increased scatter broadening associated with the same gaseous structures causing the observed DM.

We wish to compute the dispersion measure limit, $\mathrm{DM_{cutoff}}$, at which a given FRB would have been undetectable. The S/N of a detected burst depends on its intrinsic (or scatter-broadened) width, w, the time resolution of the detection system, $t_{\rm res}$, and the amount of dispersion measure time smearing between adjacent 1 MHz spectral channels, $t_{\rm smear}(\mathrm{DM})$. The resulting width of the pulse is:

$$\Delta t_{\rm obs}(\mathrm{DM}) = \sqrt{w^2 + t_{\rm res}^2 + t_{\rm smear}^2(\mathrm{DM})} \tag{9}$$

We compute $\mathrm{DM_{cutoff}}$ such that the burst, detected by our system with a signal-to-noise ratio of s_0 at a $\mathrm{DM_{obs}}$ would have fallen below our detection threshold of $s_{\rm d} = 9.0\sigma$. For each burst we thus solve

$$s_{\rm d} = s_0\sqrt{\frac{\Delta t_{\rm obs}(\mathrm{DM_{obs}})}{\Delta t_{\rm obs}(\mathrm{DM_{cutoff}})}} \tag{10}$$

Extended Data Table 1 lists the DM, widths, time resolution, detection S/N values and derived $\mathrm{DM_{cutoff}}$ values for each of the bursts in our sample.

MCMC analysis

To complement the likelihood analysis presented in the main text, we have performed Bayesian inference of a model constructed to describe the DM and redshift measurements of the FRBs. The model consists of four parameters describing two PDFs for distinct components of the dispersion measure: (i) $\mathrm{DM_{cosmic}}$, which describes the extragalactic dispersion measure including both the diffuse IGM and the gas associated with intervening galactic haloes; and (ii) $\mathrm{DM_{host}}$, which describes ionized gas associated with the host galaxy (we assume a fixed $\mathrm{DM_{MW,halo}}$ value of 50 pc cm^{-3} for the Galactic halo). We parameterize the former PDF with equation (4), that is, $p_{\rm cosmic}(\Delta)$ with $\Delta \equiv \mathrm{DM_{cosmic}}/\langle\mathrm{DM_{cosmic}}\rangle$ and $\langle\mathrm{DM_{cosmic}}\rangle$ the average value for the assumed cosmology (equation (2)). The foregoing subsection on cosmology and host-galaxy parameter estimation describes theoretical treatments that motivate one to adopt $\alpha = 3$ and $\beta = 3$ in equation (4) and to adopt the functional form of $\sigma_{\rm DM} = F/z^{1/2}$ for its dispersion parameter. For $\langle\mathrm{DM_{cosmic}}\rangle$, we modulate its amplitude via the product $\Omega_b h_{70}$. Therefore, $p_{\rm cosmic}(\Delta)$ is governed by two free parameters: F and $\Omega_b h_{70}$.

We adopt the same $p_{\rm host}(\mathrm{DM_{host}}|\mu,\sigma_{\rm host})$ PDF described earlier, with free parameters $\exp(\mu)$ and $\sigma_{\rm host}$. From these two PDFs we construct a likelihood function for the set of observed FRBs using equations (5) and (7). Note that measurement uncertainty in $\mathrm{DM_{FRB}}$ does not enter into the evaluation of $\mathcal{L}$ because the dispersion from $\mathrm{DM_{cosmic}}$ and $\mathrm{DM_{host}}$ are much greater. Put another way, our model is constructed to

describe the observed distribution of DM'_{FRB} values with an anticipated dispersion substantially exceeding the uncertainty in individual DM_{FRB} measurements (typically <1 pc cm^{-3}).

Effectively, two of the parameters ($\Omega_b h_{70}$, μ) set the amplitude of the DM–z relation and two describe its dispersion (F, σ_{host}). We anticipate a degeneracy between each set although if DM_{host} is approximately independent of redshift then this apparent degeneracy may be resolved. Only the dispersion in DM_{cosmic}, parameterized by F, allows for large negative excursions from the mean relation. Lastly, we introduce priors for the four parameters based on a combination of experimentation, physical expectation and scientific motivation. For $\Omega_b h_{70}$, the scientific focus of this manuscript, we adopt a uniform prior ranging from 0.015–0.095 which easily spans the Planck15 estimate. For F, we adopt a uniform prior in the interval (0.01, 0.5), a larger range than anticipated by our models in the frequentist analysis presented above. Regarding exp(μ), we adopt a uniform prior in the interval [20, 200] pc cm^{-3}. We consider lower values for the mean to be non-physical and we will find that larger values are disfavoured by the observations. Lastly, we assume a uniform prior for σ_{host} in the interval [0.2, 2]. The larger σ_{host} values give non-negligible probability for DM_{host} values in excess of 1,000 pc cm^{-3}. Future observations, especially an ensemble of FRBs at low redshift, will better inform these priors on μ and σ_{host}.

Adopting the above likelihood and priors, we performed a Bayesian inference of the four parameters using the gold-standard sample of FRB measurements and standard MCMC techniques. These were performed with the PYMC3 software package using slice sampling and four independent chains of 40,000 samples after a tuning period of 2,000 samples. Extended Data Fig. 5 presents a corner plot of the combined samples. A principal result is that the data yield a $\Omega_b h_{70}$ distribution fully consistent with the independent estimates from the CMB, BBN and supernovae. Quantitatively, the $\Omega_b h_{70}$ samples have a median value of 0.056 and a 68% confidence interval spanning [0.046, 0.066] (see Extended Data Table 2). Taken strictly, at 95% confidence these FRB measurements require a universe with at least 70% of the baryons inferred from BBN and CMB analysis. These results hold despite the weak priors placed on the PDF for DM_{host}, but we warn that they are dependent on the value assumed for $DM_{MW,halo}$.

Extended Data Figure 5 also reveals the anticipated anti-correlations between μ and $\Omega_b h_{70}$ and (to a lesser extent) F and σ_{host}. We expect these to weaken as the FRB sample grows in size and redshift range. Lastly, we note that the F and μ parameters have maximal probability at one edge of their assumed prior intervals. Values of F that are on the higher side of the considered range (a range that spans the possible model space) are modestly favoured. For μ, we consider 20 pc cm^{-3} to be the lowest sensible mean contribution from the host galaxy (which could also mean a lower value for the Galactic halo than adopted here).

The frequentist analysis in the main text and this Bayesian MCMC analysis agree very well on the gold-standard sample. The most notable differences are that the MCMC analysis prefers a distribution for exp(μ) that is more peaked to smaller exp(μ) values and one for F that peaks towards larger values, although with no value for exp(μ) or F strongly preferred by either analysis. When the parameters are not well constrained one would not expect perfect agreement between the methods, as, for example, the Bayesian analysis is sensitive to our prior on exp(μ) when this parameter is not well constrained. It is expected that the differences between the two methods will become smaller with more data. Already for the seven-burst sample (Extended Data Fig. 4), the distribution for F in the frequentist analysis is more similar to the MCMC analysis of the gold-standard sample.

Data availability

The data sets generated during and/or analysed during this study are available at https://data-portal.hpc.swin.edu.au/dataset/observations-of-four-localised-fast-radio-bursts-and-their-host-galaxies.

Code availability

Custom code is available at https://github.com/FRBs/FRB.

27. Mahony, E. K. et al. A search for the host galaxy of FRB 171020. *Astrophys. J.* **867**, L10 (2018).
28. Connor, L. Interpreting the distributions of FRB observables. *Mon. Not. R. Astron. Soc.* **487**, 5753–5763 (2019).
29. Schnitzeler, D. H. F. M. Modelling the Galactic distribution of free electrons. *Mon. Not. R. Astron. Soc.* **427**, 664–678 (2012).
30. Yao, J. M., Manchester, R. N. & Wang, N. A new electron-density model for estimation of pulsar and FRB distances. *Astrophys. J.* **835**, 29 (2017).
31. Oppermann, N. et al. An improved map of the Galactic Faraday sky. *Astron. Astrophys.* **542**, A93 (2012).
32. Akahori, T. & Ryu, D. Faraday rotation measure due to the intergalactic magnetic field. *Astrophys. J.* **723**, 476–481 (2010).
33. Bernet, M. L., Miniati, F., Lilly, S. J., Kronberg, P. P. & Dessauges-Zavadsky, M. Strong magnetic fields in normal galaxies at high redshift. *Nature* **454**, 302–304 (2008).
34. Ravi, V. et al. The magnetic field and turbulence of the cosmic web measured using a brilliant fast radio burst. *Science* **354**, 1249–1252 (2016).
35. Bhat, N. D. R., Cordes, J. M., Camilo, F., Nice, D. J. & Lorimer, D. R. Multifrequency observations of radio pulse broadening and constraints on interstellar electron density microstructure. *Astrophys. J.* **605**, 759–783 (2004).
36. Krishnakumar, M. A., Mitra, D., Naidu, A., Joshi, B. C. & Manoharan, P. K. Scatter broadening measurements of 124 pulsars at 327 MHz. *Astrophys. J.* **804**, 23 (2015).
37. Michilli, D. et al. An extreme magneto-ionic environment associated with the fast radio burst source FRB 121102. *Nature* **553**, 182–185 (2018).
38. Deller, A. T. et al. DiFX-2: a more flexible, efficient, robust, and powerful software correlator. *Publ. Astron. Soc. Pacif.* **123**, 275–287 (2011).
39. Kettenis, M., van Langevelde, H. J., Reynolds, C. & Cotton, B. ParselTongue: AIPS talking Python. In *Astronomical Data Analysis Software and Systems XV* (eds Gabriel, C. et al.) 497–500 (Astronomical Society of the Pacific, 2006).
40. Greisen, E. W. AIPS, the VLA, and the VLBA in *Information Handling in Astronomy — Historical Vistas* (ed. Heck, A.) 109–125 (Kluwer Academic Publishers, 2003).
41. Becker, R. H., White, R. L. & Helfand, D. J. The FIRST survey: faint images of the radio sky at twenty centimeters. *Astrophys. J.* **450**, 559 (1995).
42. Appenzeller, I. et al. Successful commissioning of FORS1 — the first optical instrument on the VLT. *Messenger* **94**, 1–6 (1998).
43. PypeIt Documentation Release 1.0.1. https://pypeit.readthedocs.io/_/downloads/en/latest/pdf/ (2020).
44. Abazajian, K. N. et al. The seventh data release of the Sloan Digital Sky Survey. *Astrophys. J. Suppl. Ser.* **182**, 543–558 (2009).
45. Prochaska, J. X. The igmspec database of public spectra probing the intergalactic medium. *Astron. Comput.* **19**, 27–33 (2017).
46. Vernet, J. et al. X-shooter, the new wide band intermediate resolution spectrograph at the ESO Very Large Telescope. *Astron. Astrophys.* **536**, A105 (2011).
47. Gimeno, G. et al. On-sky commissioning of Hamamatsu CCDs in GMOS-S. *Proc. SPIE* **9908**, 99082S (2016).
48. Freudling, W. et al. Automated data reduction workflows for astronomy. The ESO Reflex environment. *Astron. Astrophys.* **559**, A96 (2013).
49. Berriman, G. B. & Good, J. C. The application of the Montage Image Mosaic Engine to the visualization of astronomical images. *Publ. Astron. Soc. Pacif.* **129**, 058006 (2017).
50. Lang, D., Hogg, D. W., Mierle, K., Blanton, M. & Roweis, S. Astrometry.net: blind astrometric calibration of arbitrary astronomical images. *Astron. J.* **139**, 1782–1800 (2010).
51. Lindegren, L. et al. Gaia Data Release 2: the astrometric solution. *Astron. Astrophys.* **616**, A2 (2018).
52. Abbott, T. M. C. et al. The Dark Energy Survey data release 1. *Astrophys. J. Suppl. Ser.* **239**, 25 (2018).
53. Craig, M. et al. astropy/ccdproc: v1.3. 0.post1 https://doi.org/10.5281/zenodo.1069648 (2017).
54. Wright, E. A cosmology calculator for the World Wide Web. *Publ. Astron. Soc. Pacif.* **118**, 1711–1715 (2006).
55. Inoue, S. Probing the cosmic reionization history and local environment of gamma-ray bursts through radio dispersion. *Mon. Not. R. Astron. Soc.* **348**, 999–1008 (2004).
56. Deng, W. & Zhang, B. Cosmological implications of fast radio burst/gamma-ray burst associations. *Astrophys. J.* **783**, L35 (2014).
57. Madau, P. & Dickinson, M. Cosmic star-formation history. *Annu. Rev. Astron. Astrophys.* **52**, 415–486 (2014).
58. Fukugita, M. & Peebles, P. J. E. The cosmic energy inventory. *Astrophys. J.* **616**, 643–668 (2004).
59. Walters, A., Ma, Y.-Z., Sievers, J. & Weltman, A. Probing diffuse gas with fast radio bursts. *Phys. Rev. D* **100**, 103519 (2019).
60. Miralda-Escudé, J., Haehnelt, M. & Rees, M. J. Reionization of the inhomogeneous universe. *Astrophys. J.* **530**, 1–16 (2000).
61. Hafen, Z. et al. The origins of the circumgalactic medium in the FIRE simulations. *Mon. Not. R. Astron. Soc.* **488**, 1248–1272 (2019).
62. Fielding, D., Quataert, E., McCourt, M. & Thompson, T. A. The impact of star formation feedback on the circumgalactic medium. *Mon. Not. R. Astron. Soc.* **466**, 3810–3826 (2017).
63. Faucher-Giguère, C.-A., Kereš, D. & Ma, C.-P. The baryonic assembly of dark matter haloes. *Mon. Not. R. Astron. Soc.* **417**, 2982–2999 (2011).
64. Sharma, P., McCourt, M., Parrish, I. J. & Quataert, E. On the structure of hot gas in haloes: implications for the L_X-T_X relation and missing baryons. *Mon. Not. R. Astron. Soc.* **427**, 1219–1228 (2012).

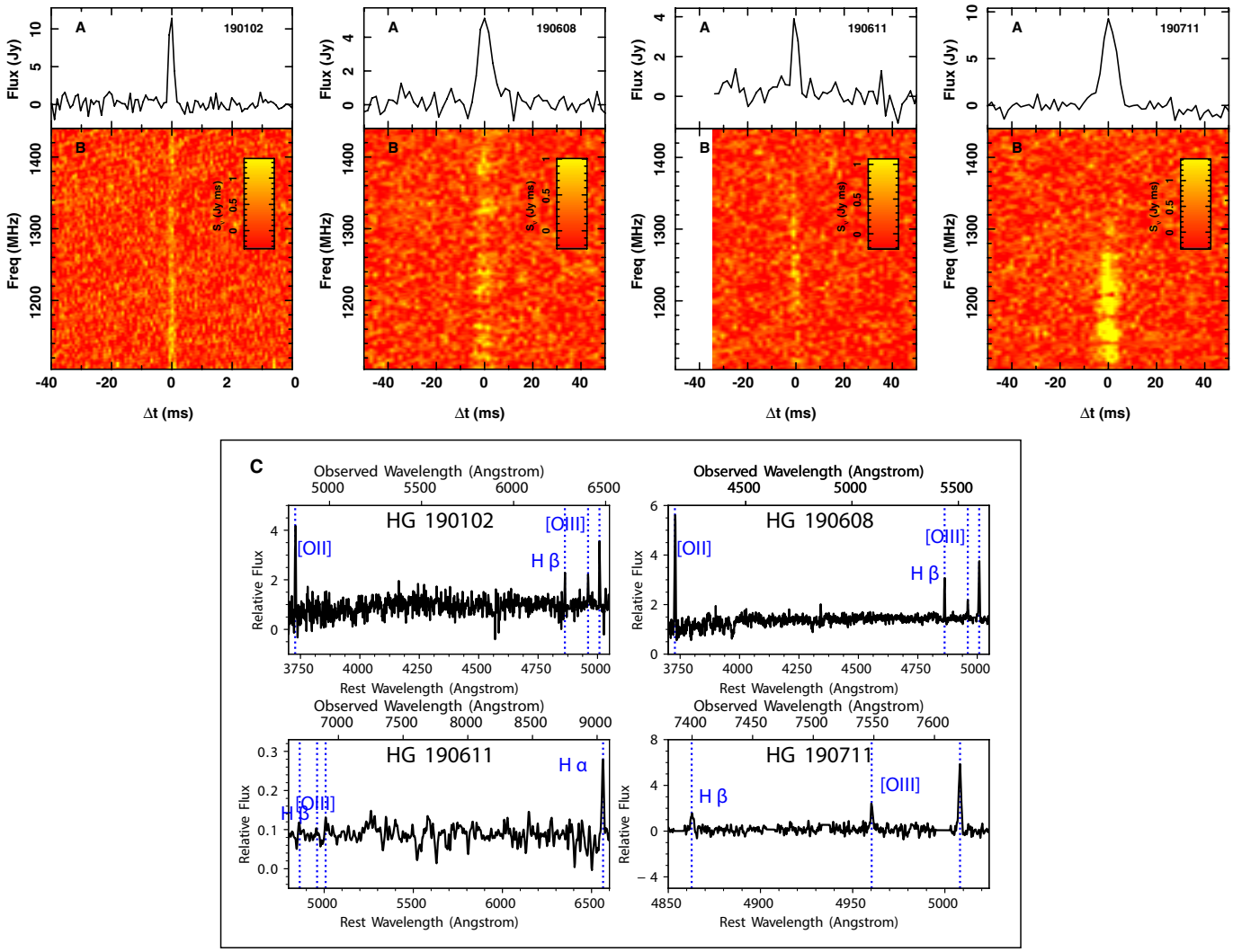
65. Voit, G. M. Ambient column densities of highly ionized oxygen in precipitation-limited circumgalactic media. *Astrophys. J.* **880**, 139 (2019).
66. Jaroszynski, M. Fast radio bursts and cosmological tests. *Mon. Not. R. Astron. Soc.* **484**, 1637–1644 (2019).
67. Debackere, S. N. B., Schaye, J. & Hoekstra, H. The impact of the observed baryon distribution in haloes on the total matter power spectrum. *Mon. Not. R. Astron. Soc.* **492**, 2285–2307 (2020).
68. Noh, Y. & McQuinn, M. A physical understanding of how reionization suppresses accretion on to dwarf haloes. *Mon. Not. R. Astron. Soc.* **444**, 503–514 (2014).
69. Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9**, 60–62 (1938); https://doi.org/10.1214/aoms/1177732360.
70. Macquart, J. P. & Ekers, R. FRB event rate counts — II. Fluence, redshift, and dispersion measure distributions. *Mon. Not. R. Astron. Soc.* **480**, 4211–4230 (2018).

Acknowledgements We thank H. Yang and L. Infante for providing IMACS imaging around FRB 190611 that informed the further follow-up observations on VLT and Gemini-S presented here. We are grateful to Australia Telescope National Facility (ATNF) operations staff and to Murchison Radio-astronomy Observatory staff for supporting our ASKAP operations, and the ATNF steering committee for allocating time for these observations. K.W.B., J.-P.M. and R.M.S. acknowledge support by Australian Research Council (ARC) grant DP180100857. A.T.D. and R.M.S. are the recipients of ARC Future Fellowships FT150100415 and FT190100155 respectively. S.O. and R.M.S. acknowledge support through ARC grant FL150100148. R.M.S. also acknowledges support through ARC grant CE170100004. N.T. acknowledges support from FONDECYT grant 11191217 and PUCV/VRIEA project 039.395/2019. The Australian Square Kilometre Array Pathfinder and Australia Telescope Compact Array are part of the Australia Telescope National Facility which is managed by CSIRO. Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Centre. Establishment of ASKAP, the Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund. Part of this work was performed on the OzSTAR national facility at Swinburne University of Technology. OzSTAR is funded by Swinburne University of Technology and the National Collaborative Research Infrastructure Strategy (NCRIS). We acknowledge the Wajarri Yamatji as the traditional owners of the Murchison Radio-astronomy Observatory site. This work includes observations collected at the European Southern Observatory under ESO programmes 0102.A-0450(A), 0103.A-0101(A) and 0103.A-1010(B). This work includes data obtained from programme GS-2019B-Q-132 at the Gemini Observatory, acquired through the Gemini Observatory Archive and processed using the Gemini PYRAF package. Gemini Observatory is operated by the Association of Universities for Research in Astronomy, Inc., under a cooperative agreement with the NSF on behalf of the Gemini partnership: the National Science Foundation (US), the National Research Council (Canada), CONICYT (Chile), Ministerio de Ciencia, Tecnología e Innovación Productiva (Argentina), Ministério da Ciência, Tecnologia e Inovação (Brazil), and Korea Astronomy and Space Science Institute (Republic of Korea). PYRAF is a product of the Space Telescope Science Institute, which is operated by AURA for NASA.

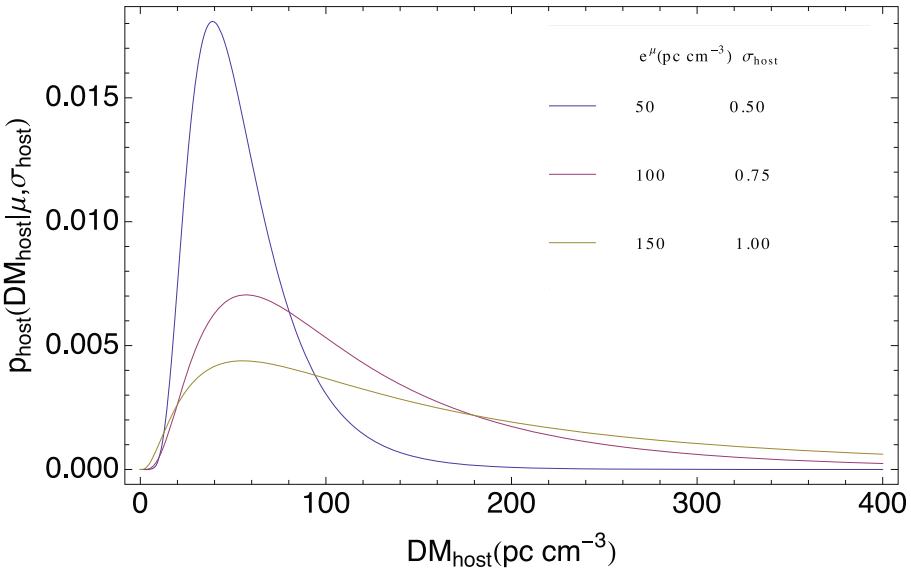
Author contributions J.X.P., M.M. and J.P.M. framed the analysis approach and drafted the manuscript, with contributions from A.T.D., R.D.E. and R.M.S. K.W.B. developed the detection and localization pipelines, and with A.T.D. and C.P. jointly developed the correlation code and interferometry processing pipeline. D.R.S. made additional improvements to the performance of the detection pipeline. R.M.S. and S.B. detected the FRBs and performed follow-up astrometry. A.T.D. and C.K.D. derived the FRB positions from the CRAFT voltage data, and S.R., J.X.P., N.T. and L.M. obtained and reduced the optical data to derive the burst host-galaxy identifications and redshifts. M.M., J.X.B. and J.-P.M. developed the IGM model and analysis code, and S.O. adapted the code to run on the Swinburne supercomputer and collated the results. C.W.J. contributed to the maximum likelihood analysis and framed the approach to estimate parameter uncertainties.

Competing interests The authors declare no competing interests.

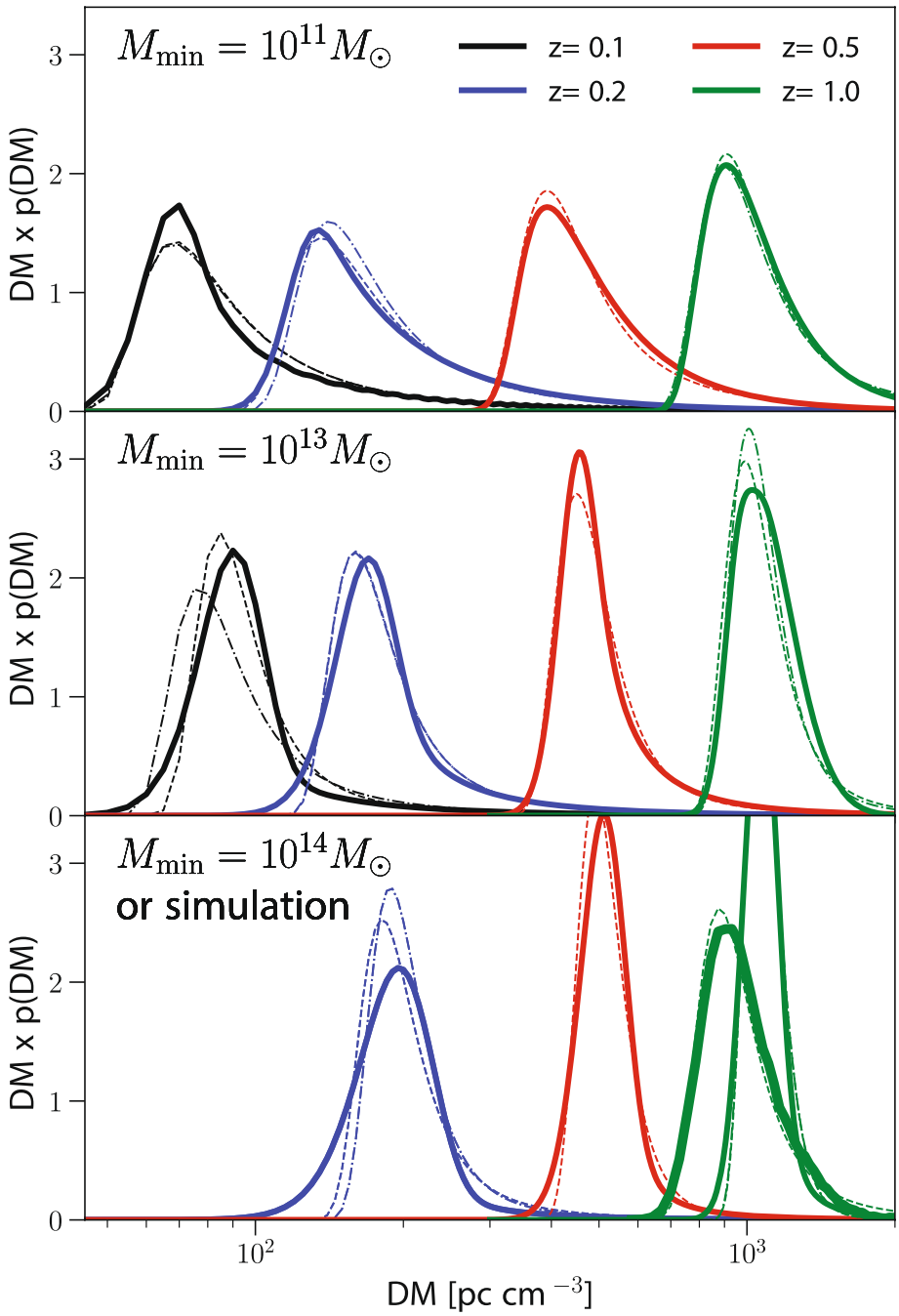
Additional information
Correspondence and requests for materials should be addressed to J.-P.M. or J.X.P.
Peer review information *Nature* thanks Kiyoshi Masu, Fabrizio Nicastro and Anthony Walters for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



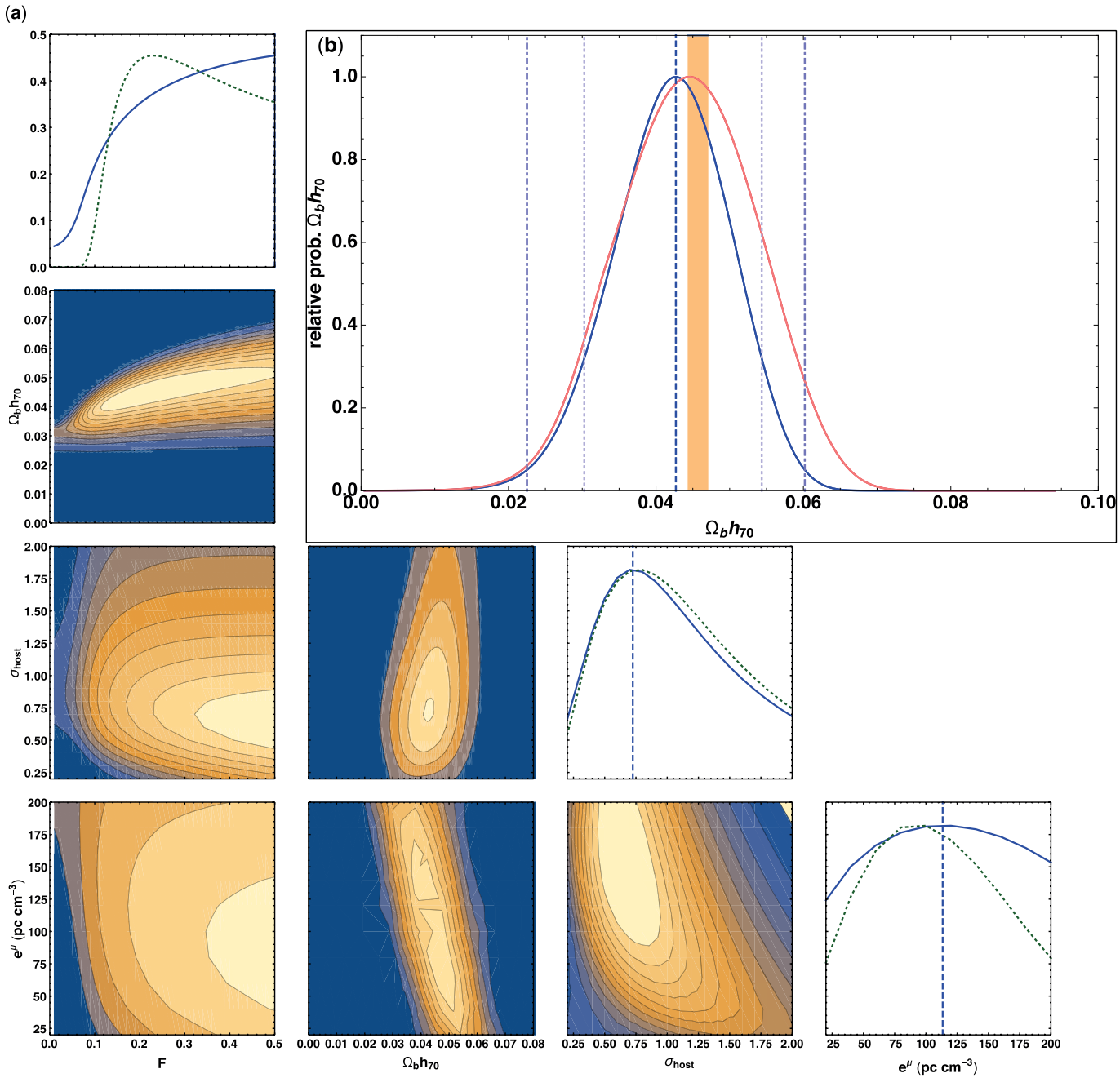
Extended Data Fig. 1 | The pulse profiles and host galaxy spectra of the four new FRBs presented here. Top row, ASKAP data. The pulse profiles (upper subpanels, labelled A) and the radio dynamic spectra (lower subpanels, labelled B) show the detections by the ASKAP incoherent capture system (ICS) of FRB 190102 with a time resolution of 0.864 ms, and of FRBs 190608, 190611 and 190711 with a resolution of 1.728 ms. The spectral resolution is 1 MHz across the 336-MHz bandwidth. Bottom row, the SDSS (HG 190608) and VLT/FORS2 (HG 190102, HG 190611 and HG 190711) optical spectra of the host galaxies located at the respective FRB positions (see Table 1), and the spectral lines from which their redshifts are deduced.



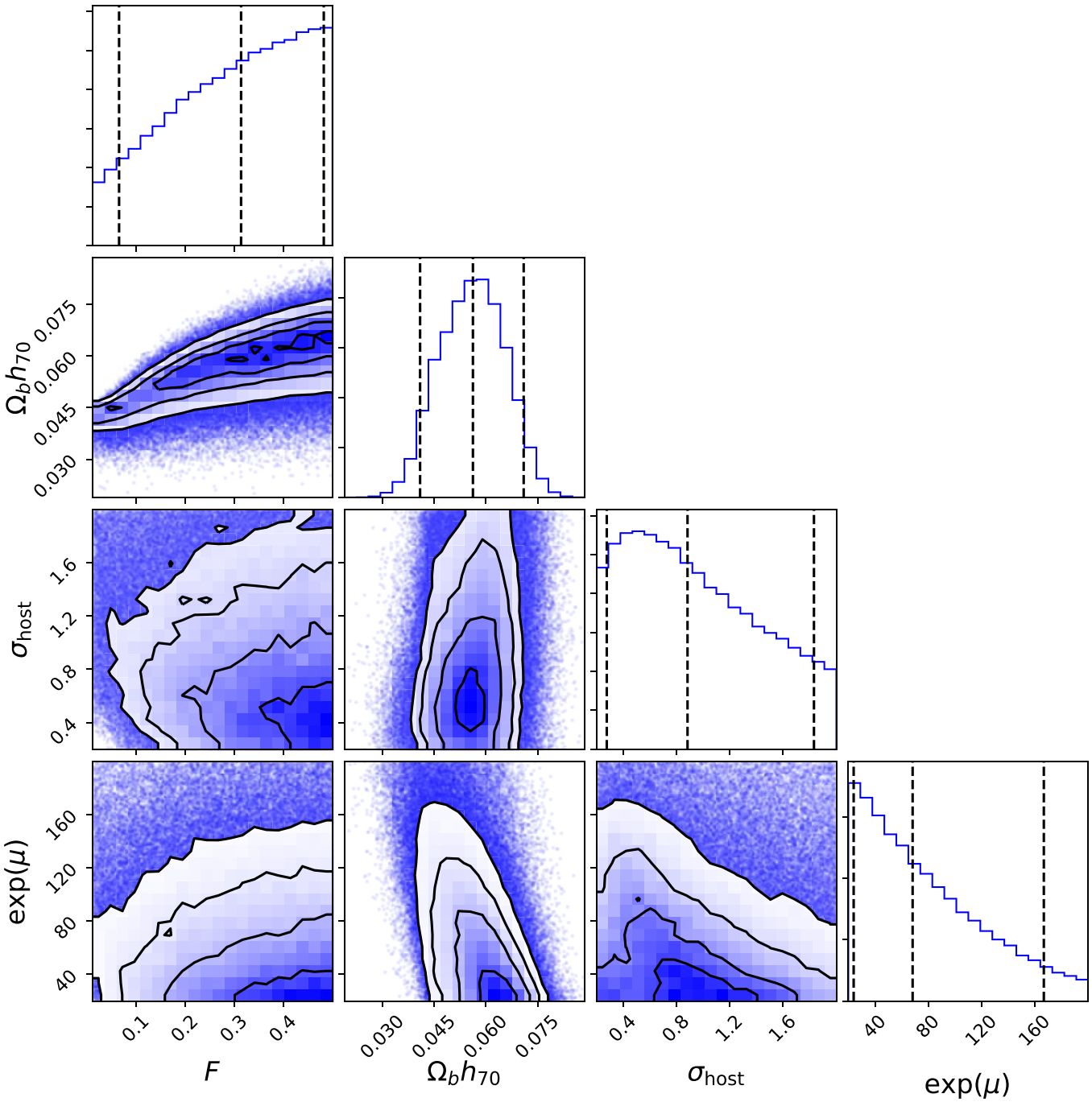
Extended Data Fig. 2 | The shape of the distribution used to model the host-galaxy dispersion measure, DM_{host}. The behaviour of the probability distribution $p_{host}(DM_{host}|\mu, \sigma_{host})$ is shown for an illustrative set of parameters spanning the range of plausible values for μ and σ_{host}.



Extended Data Fig. 3 | The expected contribution of the cosmic baryons to the dispersion measure. The probability distribution of $\mathrm{DM_{cosmic}}$ due to the cosmic baryons, p(DM), in semi-analytic models and simulations, as encoded in black, blue, red and green in order of increasing redshift (z; see key), is compared to the analytic form used in our analysis (DM; equation (4)). The thinner solid curves show semi-analytic models[11] in which the minimum halo mass that can resist feedback and retain its gas is given by $M_{\rm min}$. The dashed curves are the best-fit analytic function, and the dot-dashed curves assume the $\sigma_{\rm DM} = Fz^{-1/2}$ scaling from the $z = 0.5$ best-fit for which $F = 0.32$, 0.15 and 0.09 for the top, middle and bottom panels, respectively. Because of the success of this Euclidean-space scaling, we adopt it in our analysis. The thicker green solid curve in the bottom panel is calculated from a hydrodynamic simulation[63].



Extended Data Fig. 4 | The density of cosmic baryons derived from the extended FRB sample. The constraints on the IGM parameters $\Omega_b h_{70}$ and F, and on the host-galaxy parameters μ and σ_{host}, for a log-normal host-galaxy DM distribution are shown in an identical manner to Fig. 3, but derived using the seven-burst sample (that is, including the five gold-standard bursts as well as FRBs 190523 and 190611).



Extended Data Fig. 5 | Constraints on the cosmic baryon density and FRB host-galaxy parameters derived using a Bayesian approach. The results of a Markov Chain Monte Carlo (MCMC) analysis based on our five-FRB gold-standard sample presented in the main text demonstrate broad agreement with the results of the frequentist analysis presented in Fig. 3. The outermost vertical lines on the histogram denote the confidence region corresponding to that parameter, with the central line indicating the mean value.

Extended Data Table 1 | Detection properties of the ASKAP FRBs

FRB	detection $S/N^{(1)}$	DM ($\mathrm{pc\,cm^{-3}}$)	w ms	t_{res} ms	$\mathrm{DM_{cutoff}}$ ($\mathrm{pc\,cm^{-3}}$)
180924	21.1	361.42(6)	1.76(9)	0.864	3050
181112	19.3	589.27(3)	2.1(2)	0.864	3400
190102	14.0	363.6(3)	1.7(1)	0.864	1250
190608	16.1	338.7(5)	6.0(8)	1.728	4510
$190611^{(2)}$	9.3	321.4(2)	2(1)	1.728	430
190711	23.8	593.1(4)	6.5(5)	1.728	11500

The values of $\mathrm{DM_{cutoff}}$ denote the maximum DM at which a burst with those properties listed would have been detectable at an *S/N* threshold s_t = 9:5 with the ASKAP telescope back-end at a centre frequency of 1,295 MHz given the burst width *w* and search time resolution t_{res}, and its 1 MHz spectral resolution.

[a]The detection *S/N* value listed is that reported by the incoherent detection pipeline for the telescope beam in which the detection signal was strongest.

[b]The voltage-capture system enables the follow-up of subthreshold events detected in the incoherent pipeline, and subsequent interferometric validation, which would increase the *S/N* of a valid event by a factor $\gtrsim$5. The reported $\mathrm{DM_{cutoff}}$ is referenced to the threshold s_t = 9.0 relevant to the observing run during which this event was detected.

Extended Data Table 2 | Results of the MCMC analysis

Parameter	Unit	Prior	Median	68%	95%
F	None	$\mathcal{U}(0.011,0.5)$	0.31	0.15,0.44	0.04,0.49
$\exp(\mu)$	pc cm^{-3}	$\mathcal{U}(20.0,200)$	68.2	33.2,127.8	22.0,181.1
$\sigma_{\rm host}$	None	$\mathcal{U}(0.2,2)$	0.88	0.43,1.53	0.24,1.91
$\Omega_b h_{70}$	None	$\mathcal{U}(0.015,0.095)$	0.056	0.046,0.066	0.038,0.073

Estimates of the range of parameters from this analysis at the 68% and 95% confidence levels. These are consistent with the results of the approach described in the main text.

Article

Spectroscopy of short-lived radioactive molecules

https://doi.org/10.1038/s41586-020-2299-4

Received: 24 July 2019

Accepted: 13 March 2020

Published online: 27 May 2020

Open access

Check for updates

R. F. Garcia Ruiz[1,2]✉, R. Berger[3]✉, J. Billowes[4], C. L. Binnersley[4], M. L. Bissell[4], A. A. Breier[5], A. J. Brinson[2], K. Chrysalidis[1], T. E. Cocolios[6], B. S. Cooper[4], K. T. Flanagan[4,7], T. F. Giesen[5], R. P. de Groote[8], S. Franchoo[9], F. P. Gustafsson[6], T. A. Isaev[10], Á. Koszorús[6], G. Neyens[1,6], H. A. Perrett[4], C. M. Ricketts[4], S. Rothe[1], L. Schweikhard[11], A. R. Vernon[4], K. D. A. Wendt[12], F. Wienholtz[1,11], S. G. Wilkins[1] & X. F. Yang[13]

Molecular spectroscopy offers opportunities for the exploration of the fundamental laws of nature and the search for new particle physics beyond the standard model[1–4]. Radioactive molecules—in which one or more of the atoms possesses a radioactive nucleus—can contain heavy and deformed nuclei, offering high sensitivity for investigating parity- and time-reversal-violation effects[5,6]. Radium monofluoride, RaF, is of particular interest because it is predicted to have an electronic structure appropriate for laser cooling[6], thus paving the way for its use in high-precision spectroscopic studies. Furthermore, the effects of symmetry-violating nuclear moments are strongly enhanced[5,7–9] in molecules containing octupole-deformed radium isotopes[10,11]. However, the study of RaF has been impeded by the lack of stable isotopes of radium. Here we present an experimental approach to studying short-lived radioactive molecules, which allows us to measure molecules with lifetimes of just tens of milliseconds. Energetically low-lying electronic states were measured for different isotopically pure RaF molecules using collinear resonance ionisation at the ISOLDE ion-beam facility at CERN. Our results provide evidence of the existence of a suitable laser-cooling scheme for these molecules and represent a key step towards high-precision studies in these systems. Our findings will enable further studies of short-lived radioactive molecules for fundamental physics research.

Molecular systems provide a versatile physical environment in which to study the fundamental symmetries of nature and the interactions and properties of subatomic particles[1,2,12,13]. Among the four known fundamental forces, the weak force is the only one that is known to violate symmetry with respect to spatial inversion of all particle coordinates (known as parity violation), giving rise to various intriguing phenomena. Some of these parity-violating effects have been measured with high accuracy in atomic systems[13–15], contributing to the most stringent low-energy tests of the Standard Model of particle physics. In certain molecules, effects resulting from both parity violation (P-odd) and time-reversal violation (T-odd) are considerably enhanced with respect to atomic systems[5,7,8,13,16], offering the means to explore unknown aspects of the fundamental laws of physics. The strengths of these interactions scale with atomic number, nuclear spin and nuclear deformation, and so molecular compounds of heavy radioactive nuclei are predicted to exhibit unprecedented sensitivity, with an enhancement of more than two orders of magnitude for effects that are P-odd or simultaneously P- and T-odd[5–8,17–20].

However, the experimental knowledge of radioactive molecules is scarce[21], and quantum chemistry calculations often constitute the only source of information. Molecules possess complex quantum level structures, which renders spectroscopy of their structure considerably more challenging compared to atoms. Moreover, major additional experimental challenges must be overcome to study molecules containing heavy and deformed nuclei, which can have lifetimes of just a few milliseconds. These radioactive nuclei are very rare in nature or do not occur naturally and so must be produced artificially at specialized facilities, such as at the Isotope Separator On-line Device (ISOLDE) at CERN. Furthermore, molecules containing short-lived isotopes can only be produced in quantities smaller than 10^{-8} g (typically with rates of less than 10^{6} particles s^{-1}). Thus, spectroscopic studies require particularly sensitive experimental techniques adapted to the properties of radioactive ion beams and the conditions present at radioactive-beam facilities. Here, we present an approach for performing laser spectroscopy of short-lived radioactive molecules, using the highly sensitive collinear resonance ionization method[22]. These results provide the first

[1]CERN, Geneva, Switzerland. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Fachbereich Chemie, Philipps-Universität Marburg, Marburg, Germany. [4]Department of Physics and Astronomy, The University of Manchester, Manchester, UK. [5]Laboratory for Astrophysics, Institute of Physics, University of Kassel, Kassel, Germany. [6]KU Leuven, Instituut voor Kern- en Stralingsfysica, Leuven, Belgium. [7]Photon Science Institute, The University of Manchester, Manchester, UK. [8]Department of Physics, University of Jyväskylä, Jyväskylä, Finland. [9]Institut de Physique Nucleaire d'Orsay, Orsay, France. [10]NRC 'Kurchatov Institute'-PNPI, Gatchina, Russia. [11]Institut für Physik, Universität Greifswald, Greifswald, Germany. [12]Institut für Physik, Johannes Gutenberg-Universität Mainz, Mainz, Germany. [13]School of Physics and State Key Laboratory of Nuclear Physics and Technology, Peking University, Beijing, China. ✉e-mail: rgarciar@mit.edu; robert.berger@uni-marburg.de

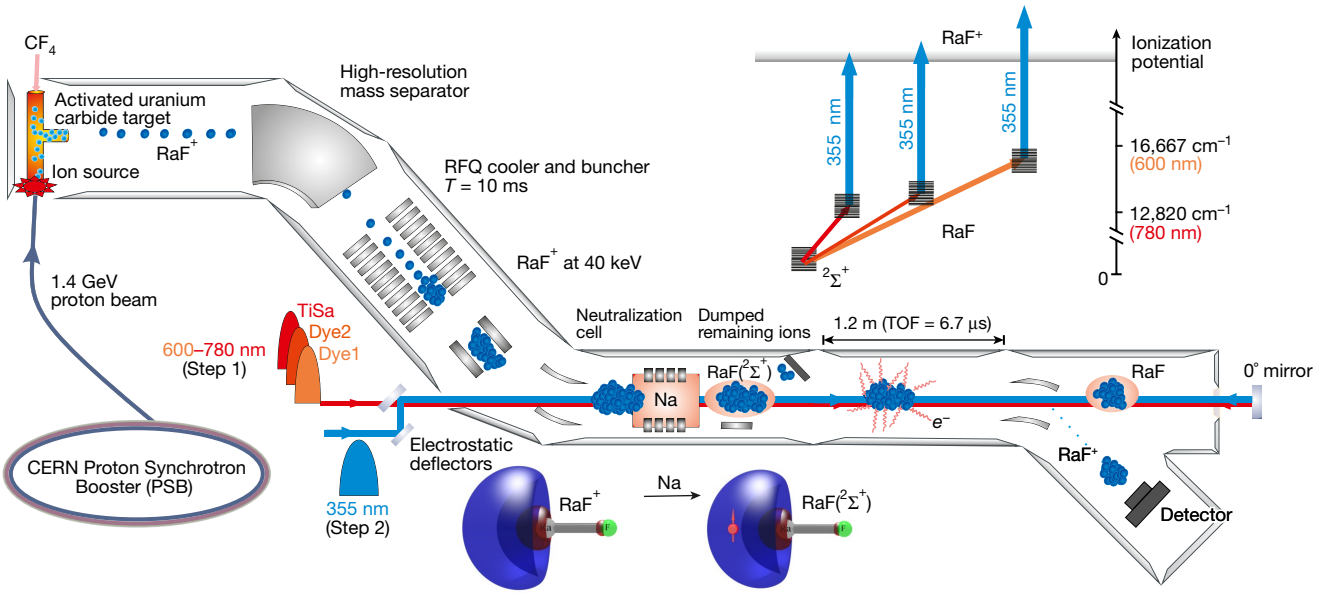


Fig. 1 | Experimental scheme for the production and study of short-lived radioactive molecules. Radioactive radium isotopes were created by impinging 1.4-GeV protons from the CERN Proton Synchrotron Booster (PSB) on a uranium carbide (UC_x) target. Radium monofluoride cations (RaF^+) were produced by passing tetrafluoromethane (CF_4) gas through the activated UC_x target at 1,300 °C. Molecular ions were extracted from the source, mass-selected and injected into a helium-filled RFQ trap, where they were accumulated for 10 ms. Bunches of molecular ions were extracted and neutralized in flight by charge exchange with neutral sodium atoms. Neutral RaF molecules were overlapped with different laser beams (step 1, TiSa, Dye1 and Dye2, and step 2, a 355-nm laser; see Methods section 'Laser setup') in a collinear geometry. Resonantly reionized molecules were deflected onto a particle detector. The resonance ionization scheme is shown at top right. At bottom, molecular orbitals are shown schematically. Nuclear positions within the molecules are coarsely indicated by a grey sphere (Ra) and green sphere (F), and the sigma bond between the atoms is indicated by the grey cylinders. Further details are provided in 'Experimental scheme'.

spectroscopic information of RaF, including isotopologues composed of radioactive isotopes with lifetimes as short as a few days. To our knowledge, this is the first laser spectroscopy study performed on a molecule containing a short-lived isotope. Moreover, this experimental scheme can be applied to study other radioactive molecules, even those composed of isotopes with lifetimes as short as a few tens of milliseconds.

Since the direct cooling of diatomic molecules with lasers[23] was experimentally demonstrated[24], there has been a wealth of studies on laser-cooling techniques and applications in molecular physics[25–31]. In contrast to other heavy-atom molecules, RaF is predicted to have highly closed excitation and re-emission optical cycles, which would make it ideal for laser cooling and trapping[6]. Moreover, owing to the recently discovered pear-shaped nuclear deformation of certain radium isotopes[11], the interactions of the electrons with the P-odd nuclear anapole moment as well as with the P,T-odd nuclear Schiff and magnetic quadrupole moments are predicted to be enhanced by more than two orders of magnitude[4,5,19,32]. Hence, these molecules could provide a unique environment in which to measure these symmetry-violating nuclear moments.

Experimental scheme

Figure 1 shows a diagram of the experimental setup used to produce and study the RaF molecules. As a first step, radium isotopes were produced by diffusion out of an irradiated target (see Methods section 'Production of RaF molecules'). RaF^+ molecular ions were formed upon injection of CF_4 gas into the target environment. The molecular ions were extracted from the ion source by applying an electrostatic field, and molecules containing one specific radium isotope were selected with a high-resolution magnetic mass separator ($\Delta m/m \approx 1/2{,}000$). The ions were collisionally cooled in a radio-frequency quadrupole (RFQ) trap filled with helium gas at room temperature (about 300 K). After up to 10 ms of cooling time, bunches of RaF^+ with a 4-μs temporal width were released and accelerated to 39,998(1) eV, before entering into the Collinear Resonance Ionisation Spectroscopy (CRIS) setup[22,33,34]. At the CRIS beam line, the ions were first neutralized in-flight by passing through a collision cell filled with a sodium vapour, inducing charge exchange according to the reaction $RaF^+ + Na \rightarrow RaF + Na^+$. As the ionization energy of RaF is estimated to be close to that of sodium (5.14 eV)[35], the neutralization reaction dominantly populates the RaF $X^2\Sigma^+$ electronic ground state. Molecular pseudo-orbitals obtained from one-component open-shell (neutral) or closed-shell (ion) restricted Hartree–Fock calculations with an energy-consistent effective core potential on radium are shown schematically in Fig. 1 (bottom). The lowest unoccupied molecular orbital in RaF^+, which is mainly of non-bonding character, becomes occupied by an unpaired electron (symbolized in Fig. 1 by a red sphere together with an arrow representing the electron spin) upon neutralization. This is shown schematically as an isodensity, with lobes in slightly transparent blue and transparent red indicating different relative phases of the single-electron wavefunction.

After the charge-exchange reaction, non-neutralized RaF^+ ions were deflected out of the beam, and the remaining bunch of neutral RaF molecules was overlapped in time and space by several (pulsed) laser beams in a collinear arrangement, along the ultrahigh-vacuum (10^{-10} mbar) interaction region of 1.2-m length. Laser pulses (step 1) of tunable wavelength were used to resonantly excite the transition of interest, and a high-power 355-nm laser pulse (step 2) was used to subsequently ionize the excited RaF molecules into RaF^+ (see Fig. 1, top). The resonantly ionized molecules were then separated from the non-ionized molecules by deflecting the ions onto a particle detector. When the excitation laser is on resonance with a transition in the molecule (step 1 in Fig. 1), the second laser pulse ionizes the molecule, producing a signal at the detector. Molecular excitation spectra were obtained by monitoring the ion counts as a function of the wavenumber of the first laser.

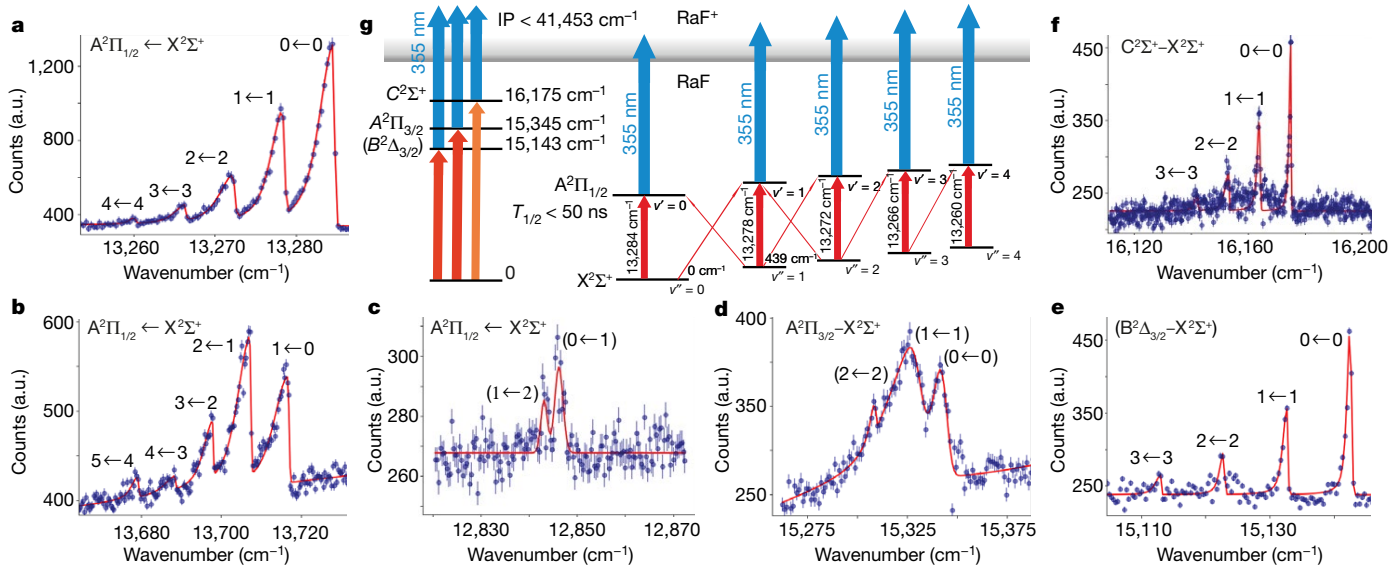


Fig. 2 | Examples of vibronic spectra measured for ^{226}RaF. a–f, The counts on the particle detector were measured as a function of the laser wavenumber of the resonant step. A fixed wavelength (355 nm) was used for the ionization step. **a**, The observed peaks corresponding to the vibronic spectra of the $\Delta v = 0$ band system of $v'' = 0, 1, 2, 3, 4$, scanned by the grating Ti:sapphire laser. **b**, **c**, The pulsed dye laser was used to scan electronic transitions in different wavelength ranges: the $\Delta v = +1$ band system of the $A^2\Pi_{1/2} \leftarrow X^2\Sigma^+$ transition with $v'' = 0, 1, 2, 3, 4$ (**b**) and the $(v', v'') = (0, 1)$ and $(1, 2)$ band. **d**–**f**, The corresponding transitions to other electronic states: $A^2\Pi_{3/2} \leftarrow X^2\Sigma^+$ (**d**), $B^2\Delta_{3/2} \leftarrow X^2\Sigma^+$ (tentatively assigned; **e**) and $C^2\Sigma^+ \leftarrow X^2\Sigma^+$ (**f**). The shape of the spectra is due to population distribution of different rotational states. The solid lines show the fit with skewed Voigt profiles. **g**, Scheme of the molecular energy levels. The estimated upper limit of the ionization potential (IP) is indicated. Three essential properties for laser cooling of RaF molecules were identified: 1) the short lifetime of the excited states $^2\Pi_{1/2}$ ($T_{1/2} < 50$ ns), which will allow for the application of strong optical forces; 2) dominant diagonal transitions, $(\Delta v = 0)/(\Delta v = \pm 1, \Delta v = 0) > 0.97$, indicating a large diagonal Franck–Condon factor; and 3) the expected low-lying electronic states $B^2\Delta_{3/2}$, $A^2\Pi_{3/2}$ and $C^2\Sigma^+$ were found to be above the $A^2\Pi_{1/2}$ states, which will enable efficient optical-cooling cycles. Wavenumbers in the spectra are given in the rest frame of the molecule. In **a**–**f**, the error bars show the statistical uncertainties (1 standard deviation) for the number of resonantly ionized molecules obtained within each laser frequency interval.

Only theoretical predictions were available for the excitation energies of RaF, and so finding the transition experimentally required scanning a large wavelength range (>1,000 cm^{-1}). The prediction for the $A^2\Pi_{1/2}$–$X^2\Sigma^+$ (0, 0) transition, for example, was 13,300 cm^{-1}, with an accuracy estimated to be within 1,200 cm^{-1} (refs. [6,32]). Given the bandwidths of the commonly available lasers (<0.3 cm^{-1}), the scan of such a large wavelength region on samples produced at rates below 10^6 molecules s^{-1} represented a major experimental challenge. To optimize the search of molecular transitions, three broadband lasers were scanned simultaneously and both collinearly and anti-collinearly (see Methods section 'Laser setup').

Results

The predicted region for the $A^2\Pi_{1/2} \leftarrow X^2\Sigma^+$ transition was scanned at a speed of 0.06 cm^{-1} s^{-1}, covering a range of 1,000 cm^{-1} in about 5 h, using the six simultaneously applied scanning regions. After a few hours of scanning on a beam of ^{226}RaF, a clear sequence of vibronic absorption signals was recorded. The measured spectrum assigned to the (v', v'') vibrational transitions (0, 0), (1, 1), (2, 2), (3, 3) and (4, 4) of the $A^2\Pi_{1/2}$–$X^2\Sigma^+$ band system is shown in Fig. 2a. Weaker band structures, that were found at about +440 cm^{-1} and −440 cm^{-1} with respect to the (0, 0) band, were assigned to the $\Delta v = \pm 1$ transitions $(v', v'') = (1, 0)$, (2, 1), (3, 2), (4, 3), (5, 4) and $(v', v'') = (0, 1)$, (1, 2), respectively (Fig. 2b, c). The quantum number assignment for $\Delta v = -1$ is tentative, owing to the highly dense structure of overlapping vibronic bands.

In addition to the $A^2\Pi_{1/2}$–$X^2\Sigma^+$ band system, we found spectroscopic signatures of electronic transitions to higher-lying states. Some examples of recorded spectra are shown in Fig. 2d–f, along with the energy-level scheme. We assign the observed transitions as follows: 1) The band system around 15,325 cm^{-1} (Fig. 2d) is attributed to the $A^2\Pi_{3/2}$–$X^2\Sigma^+$ transition, owing to the complex rovibrational structure expected to arise from the intense satellites that are possible in these transitions. Because the bands are comparatively strong, they are assigned to the $\Delta v = 0$ band system. Although the individual assignments to vibrational transitions must be considered to be tentative, as per the congested structure of the Franck–Condon profile, the $\Delta v = 0$ assignment is substantiated because no additional structure was located within a relative range of −400 to +400 cm^{-1}. The band system located around 15,143 cm^{-1} (Fig. 2e) is tentatively assigned to the $B^2\Delta_{3/2}$–$X^2\Sigma^+$ transition by virtue of the good agreement with the computed excitation energies to the $\Omega = 3/2$ state of mixed Δ/Π character[6,32]. This mixing provides intensity to the one-photon transition from a Σ state into the Δ manifold. The computed Born–Oppenheimer potentials for this $\Omega = 3/2$ state and the electronic ground state are, however, highly parallel, which would suggest a sparser Franck–Condon profile than was observed experimentally. However, we note that the related $B^2\Delta_{3/2}$–$X^2\Sigma^+$ transition in BaH and BaD was reported to have a perturbed character owing to mixing between electronic levels[36]. Thus, in the present case, a vibrational profile that is richer than expected from adiabatic potentials cannot be ruled out a priori. The band system with origin at 16,175 cm^{-1} (Fig. 2f) is assigned to the $C^2\Sigma^+$–$X^2\Sigma^+$ transition on the basis of the observed Franck–Condon profile, which is in good agreement with the computed harmonic vibrational energy spacings as well as the expected intensity distribution, and is in a wavenumber region that is only slightly lower than predicted[6,32]. All measured and assigned vibronic bands of the four electronic transitions are listed in Table 1.

The measured $A^2\Pi_{1/2}$–$X^2\Sigma^+$ (0, 0) band centre, $\tilde{T}_e = 13{,}287.8(1)$ cm^{-1} is in excellent agreement with the ab initio calculated value of 13,300(1,200) cm^{-1} (ref. [32]). In accordance with theoretical predictions[6], we found vibronic transitions with $\Delta v = 0$ to be much stronger than those with $\Delta v = \pm 1$. For most of the measurements, the power density used for the resonant step was 100(5) μJ cm^{-2} per pulse, as measured

Table 1 | Measured vibronic transitions of ^{226}RaF from the $X^2\Sigma^+$ electronic ground state to the excited $A^2\Pi$ and $B^2\Delta$ states

Transition	$v' \leftarrow v''$	$\Delta\tilde{\nu}$ (cm^{-1})
$A^2\Pi_{1/2} \leftarrow X^2\Sigma^+$	0–0	13,284.7(5)
	1–1	13,278.5(5)
	2–2	13,272.4(5)
	3–3	13,266.4(10)
	4–4	13,260.2(10)
	1–0	13,716.9(5)
	2–1	13,707.4(5)
	3–2	13,698.0(5)
	4–3	13,688.6(10)
	5–4	13,679.4(10)
	(0–1)	12,846.3(10)
	(1–2)	12,843.1(10)
($B^2\Delta_{3/2} \leftarrow X^2\Sigma^+$)	0–0	15,142.7(5)
	1–1	15,132.8(10)
	2–2	15,123.0(10)
	3–3	15,113.2(10)
$A^2\Pi_{3/2} \leftarrow X^2\Sigma^+$	(0–0)	15,344.6(50)
	(1–1)	15,325.0(80)
	(2–2)	15,309.4(100)
$C^2\Sigma^+ \leftarrow X^2\Sigma^+$	0–0	16,175.2(5)
	1–1	16,164.2(5)
	2–2	16,153.4(5)
	3–3	16,142.4(10)

The values indicate the band head positions.
Combined statistical and systematic uncertainties are included in parentheses.
The $B^2\Delta_{3/2} \leftarrow X^2\Sigma^+$ assignment is tentative.

at the entry window of the beam line. Reducing the power by 50% did not reduce the resonant ionization rate, indicating that these transitions were measured well above saturation. The much weaker vibrational transitions with $\Delta v = \pm 1$ were scanned with a pulsed dye laser of 500(5) μJ cm^{-2} power density per pulse (bandwidth of 0.1 cm^{-1}). The $\Delta v = \pm 1$ transitions were measured well above saturation and with laser beams of different characteristics, and so a precise estimation of the Franck–Condon factors could not be obtained. Instead, a lower limit of 0.97 for the peak intensity ratio $I(0,0)/I(0,1)$ was derived, indicating highly diagonal Franck–Condon factors, an essential property for laser cooling[6].

By measuring the resonant ionization rate for different time delays between the excitation and ionization laser pulses, we obtained an upper limit for the lifetime of the excited state $^2\Pi_{1/2}(v' = 0)$: $T_{1/2} \leq 50$ ns. The measurements were performed with the wavenumber of the resonant laser fixed at the resonance value of the transition $(v', v'') = (0, 0)$. The resonant ionization rate dropped by more than 70% for delays above 50 ns. This short lifetime corresponds to a large spontaneous decay rate (>2×10^7 s^{-1}), which would allow for the application of strong optical forces for laser cooling. An additional concern for the suitability of laser cooling is related to the existence of metastable states lying energetically below the $^2\Pi_{1/2}$ level, which could prevent the application of a closed optical-cooling loop, a major problem encountered for BaF (ref[6].). In contrast to BaF, all other predicted electronic states ($^2\Pi_{3/2}$, $^2\Delta_{3/2}$ and $^2\Sigma$) in RaF were found to be energetically above the $^2\Pi_{1/2}$ state, indicating that its electronic structure will allow for efficient optical-cooling cycles.

From combination differences of energetically low-lying vibronic transitions in the band system $A^2\Pi_{1/2}$–$X^2\Sigma^+$, we have derived

Table 2 | ^{226}RaF Morse potential parameters for $X^2\Sigma^+$ electronic ground and $A^2\Pi_{1/2}$ excited states

Parameter	$\tilde{\omega}_e$ (cm^{-1})	$\tilde{\mathcal{D}}_e$ ($\times 10^{-4}$ cm^{-1})
$X^2\Sigma^+$	441.8(1)	2.92(5)
$A^2\Pi_{1/2}$	435.5(1)	2.90(3)

experimental values for the harmonic frequency, $\tilde{\omega}_e$, and the dissociation energies, $\tilde{\mathcal{D}}_e$, using a Morse potential approximation. Results are given in Table 2, and further details of the analysis can be found in Methods section 'Spectroscopic analysis'.

Furthermore, we measured the $A^2\Pi_{1/2} \leftarrow X^2\Sigma^+$ vibronic spectra of ^{226}RaF and the short-lived isotopologues ^{223}RaF, ^{224}RaF, ^{225}RaF, and ^{228}RaF (Fig. 3). All vibrational transitions were clearly observed, including those of the molecule with the shortest-lived radium isotope studied, ^{224}RaF ($T_{1/2} = 3.6$ d). An on-line irradiation of the target material will enable the study of molecules containing isotopes with lifetimes as short as a few tens of milliseconds. The main limitation is dictated by the release from the target and the time spent in the RFQ trap (>5 ms). Future high-resolution measurements will enable studies of nuclear structure changes resulting from different isotopes and nuclear spins.

Conclusions and future perspectives

In summary, this Article presents an experimental approach for performing laser spectroscopy studies of molecules containing radioactive nuclei, which are typically produced at rates lower than 10^6 molecules s^{-1}. Our results have established the energetically low-lying electronic structure of RaF, providing experimental evidence for the suitability of this diatomic molecule in a laser-cooling scheme. These findings are a pivotal step towards precision measurements in this system, which are expected to provide a highly sensitive environment for the exploration of physics beyond the Standard Model of particle physics.

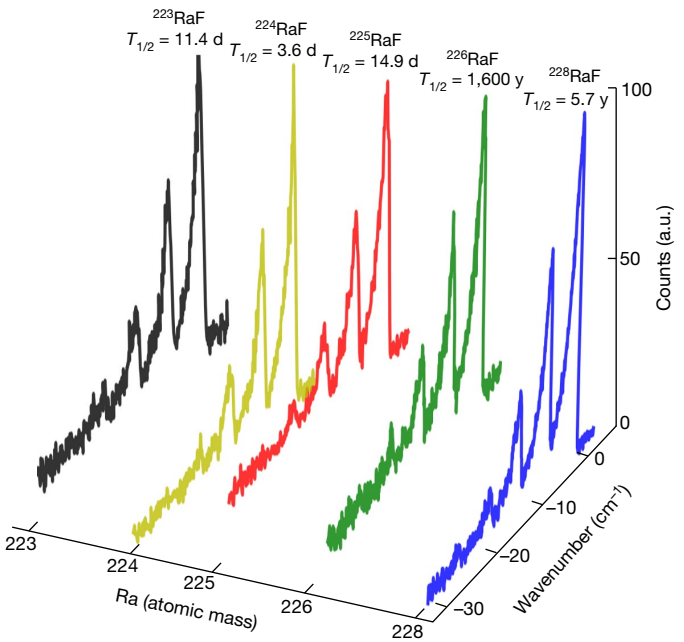


Fig. 3 | Vibronic spectra measured for different isotopologues of RaF. Measured vibronic absorption spectra for the $A^2\Pi_{1/2} \leftarrow X^2\Sigma^+$ transition are shown for the isotopologues ^{223}RaF, ^{224}RaF, ^{225}RaF, ^{226}RaF and ^{228}RaF. Wavenumber values are relative to the transition (0, 0) of ^{226}RaF.

Our experimental scheme can also be used to perform laser spectroscopy of a wide variety of neutral molecules and molecular ions, including those composed of isotopes with lifetimes of a few tens of milliseconds. Radioactive molecules can be precisely tailored to enhance their sensitivity to parity- and time-reversal-violating effects by introducing heavy and octupole-deformed nuclei. Moreover, by systematically replacing their constituent nuclei with different isotopes of the same element, both nuclear-spin-independent and nuclear-spin-dependent effects can be comprehensively studied. In addition, the present technique is applicable to other molecules of interest in studies of fundamental physics that are as yet experimentally unexplored, such as RaOH (ref. [37]), RaO (ref. [18]), RaH (ref. [17]), AcF (ref. [38]) and ^{229}ThO (ref. [5]).

In addition to the impact of our findings on quantum chemistry, nuclear structure and fundamental physics research, the ability to produce, mass-select and spectroscopically study short-lived radioactive molecules is of importance to other fields of research such as radiochemistry[21] and astrophysics[39,40]. Laboratory measurements of the spectra of radioactive molecules of astrophysical interest will allow their unambiguous identification in future astronomical observations. Furthermore, the possibility of performing spectroscopy on fast molecular beams will enable sub-Doppler spectroscopy to be performed even on molecules created at high temperatures (>600 K). Thus, we expect our results will motivate further avenues of research at the increasingly capable radioactive-ion-beam facilities around the world.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2299-4.

1. Altuntaş, E. et al. Demonstration of a sensitive method to measure nuclear-spin-dependent parity violation. *Phys. Rev. Lett*. **120**, 142501 (2018).
2. ACME Collaboration. Improved limit on the electric dipole moment of the electron. *Nature* **562**, 355–360 (2018).
3. Berger, R. & Stohner, J. Parity violation. *WIREs Comput. Mol. Sci*. **9**, e1396 (2019).
4. Flambaum, V. V., DeMille, D. & Kozlov, M. G. Time-reversal symmetry violation in molecules induced by nuclear magnetic quadrupole moments. *Phys. Rev. Lett*. **113**, 103003 (2014).
5. Flambaum, V. V. Enhanced nuclear Schiff moment and time reversal violation in ^{229}Th-containing molecules. *Phys. Rev. C* **99**, 035501 (2019).
6. Isaev, T. A., Hoekstra, S. & Berger, R. Laser-cooled RaF as a promising candidate to measure molecular parity violation. *Phys. Rev. A* **82**, 052521 (2010).
7. Auerbach, N., Flambaum, V. V. & Spevak, V. Collective T- and P-odd electromagnetic moments in nuclei with octupole deformations. *Phys. Rev. Lett*. **76**, 4316–4319 (1996).
8. Kudashov, A. D. et al. Ab initio study of radium monofluoride (RaF) as a candidate to search for parity- and time-and-parity-violation effects. *Phys. Rev. A* **90**, 052513 (2014).
9. Parker, R. H. et al. First measurement of the atomic electric dipole moment of ^{225}Ra. *Phys. Rev. Lett*. **114**, 233002 (2015).
10. Butler, P. A. et al. Evolution of octupole deformation in radium nuclei from Coulomb excitation of radioactive ^{222}Ra and ^{228}Ra beams. *Phys. Rev. Lett*. **124**, 042503 (2020).
11. Gaffney, L. et al. Studies of pear-shaped nuclei using accelerated radioactive beams. *Nature* **497**, 199–204 (2013).
12. Cairncross, W. B. et al. Precision measurement of the electron's electric dipole moment using trapped molecular ions. *Phys. Rev. Lett*. **119**, 153001 (2017).
13. Safronova, M. S. et al. Search for new physics with atoms and molecules. *Rev. Mod. Phys*. **90**, 025008 (2018).
14. Antypas, D. et al. Isotopic variation of parity violation in atomic ytterbium. *Nat. Phys*. **15**, 120–123 (2019).
15. Wood, C. S. et al. Measurement of parity nonconservation and an anapole moment in cesium. *Science* **275**, 1759–1763 (1997).
16. Sandars, P. G. H. Measurability of the proton electric dipole moment. *Phys. Rev. Lett*. **19**, 1396–1398 (1967).
17. Fazil, N. M., Prasannaa, V. S., Latha, K. V. P., Abe, M. & Das, B. P. RaH as a potential candidate for electron electric-dipole-moment searches. *Phys. Rev. A* **99**, 052502 (2019).
18. Flambaum, V. V. Electric dipole moments of actinide atoms and RaO molecule. *Phys. Rev. A* **77**, 024501 (2008).
19. Gaul, K., Marquardt, S., Isaev, T. A. & Berger, R. Systematic study of relativistic and chemical enhancements of P,T-odd effects in polar diatomic radicals. *Phys. Rev. A* **99**, 032509 (2019).
20. Sasmal, S., Pathak, H., Nayak, M. K., Vaval, N. & Pal, S. Relativistic coupled-cluster study of RaF as a candidate for the parity- and time-reversal-violating interaction. *Phys. Rev. A* **93**, 062506 (2016).
21. Formanuik, A. et al. Actinide covalency measured by pulsed electron paramagnetic resonance spectroscopy. *Nat. Chem*. **9**, 578–583 (2017).
22. Flanagan, K. T. et al. Collinear resonance ionization spectroscopy of neutron-deficient francium isotopes. *Phys. Rev. Lett*. **111**, 212501 (2013).
23. Di Rosa, M. D. Laser-cooling molecules. *Eur. Phys. J. D* **31**, 395–402 (2004).
24. Shuman, E. S. et al. Laser cooling of a diatomic molecule. *Nature* **467**, 820–823 (2010).
25. Anderegg, L. et al. Laser cooling of optically trapped molecules. *Nat. Phys*. **14**, 890–893 (2018).
26. Barry, J. F. et al. Magneto-optical trapping of a diatomic molecule. *Nature* **512**, 286–289 (2014).
27. Hudson, J. J. et al. Improved measurement of the shape of the electron. *Nature* **473**, 493–496 (2011).
28. Isaev, T. A. & Berger, R. Polyatomic candidates for cooling of molecules with lasers from simple theoretical concepts. *Phys. Rev. Lett*. **116**, 063006 (2016).
29. Isaev, T. A. & Berger, R. Towards ultracold chiral molecules. *Chimia* **72**, 375–378 (2018).
30. Lim, J. et al. Laser cooled YbF molecules for measuring the electron's electric dipole moment. *Phys. Rev. Lett*. **120**, 123201 (2018).
31. Truppe, S. et al. Molecules cooled below the Doppler limit. *Nat. Phys*. **13**, 1173–1176 (2017).
32. Isaev, T. A. & Berger, R. Lasercooled radium monofluoride: a molecular all-in-one probe for new physics. Preprint at http://arxiv.org/abs/1302.5682 (2013).
33. de Groote, R. P. et al. Use of a continuous wave laser and Pockels cell for sensitive high-resolution collinear resonance ionization spectroscopy. *Phys. Rev. Lett*. **115**, 132501 (2015).
34. Garcia Ruiz, R. F. et al. High-precision multiphoton ionization of accelerated laser-ablated species. *Phys. Rev. X* **8**, 041005 (2018).
35. Isaev, T. A. et al. Ion neutralisation mass-spectrometry route to radium monofluoride. Preprint at http://arxiv.org/abs/1310.1511 (2013).
36. Bernard, A. et al. The 5*d* states of barium hydride; BaH and BaD. *Mol. Phys*. **67**, 1–18 (1989).
37. Isaev, T. A., Zaitsevskii, A. V. & Eliav, E. Laser-coolable polyatomic molecules with heavy nuclei. *J. Phys. B* **50**, 225101 (2017).
38. Flambaum, V. V. & Dzuba, V. A. Electric dipole moments of atoms and molecules produced by enhanced nuclear Schiff moments. *Phys. Rev. A* **101**, 042504 (2020).
39. Breier, A. A. et al. Lowest bending mode of ^{13}C-substituted C_3 and an experimentally derived structure. *J. Chem. Phys*. **145**, 234302 (2016).
40. Kamiński, T. et al. Astronomical detection of radioactive molecule ^{26}AlF in the remnant of an ancient explosion. *New Astron*. **2**, 778–783 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Methods

Production of RaF molecules

Ra isotopes were produced 33 d before the laser-spectroscopy measurements by impinging 1.4 GeV protons on the cold UC_x target material. The target was exposed to pulses of 10^{13} protons per pulse over a period of 2 d. After irradiation with a total of 8×10^{17} protons, the target was kept in a sealed chamber filled with Ar gas. After day 33, the target was connected to the High-Resolution Separator (HRS) front-end at ISOLDE. FLUKA[41] simulations predicted 2×10^{13} atoms of ^{226}Ra in the target material (7.5×10^{-9} g), following proton irradiation of a cold target. The target was pumped down to pressures below 10^{-5} mbar, and the target holder and ion source were gradually heated up to about 1,300 °C, in order for the Ra isotopes to diffuse towards the surface of the target material. A leak valve attached to the target was used to inject CF_4 into the target environment. The CF_4 molecules dissociate and react with atoms and molecules on the target surface until an equilibrium is reached. RaF molecules were formed by reactive collisions of CF_4 molecules with Ra atoms present inside the irradiated target material.

According to thermodynamic equilibrium calculations[42], RaF_2 or RaF are expected to form, depending on the local temperature. Within the temperature gradient between the target (1,300 °C) and the ion source (2,000 °C), RaF_2 fully reacts to form RaF. A measured ratio of the ion-beam intensity of Ra^+ to RaF^+ of less than 0.05 indicates that more than 95% of the Ra isotopes released from the target material are converted and extracted as molecules.

The $^{226}RaF^+$ ($A = 245$) beam extracted from the ISOLDE target unit was sent to the ISOLTRAP setup[43], where the molecular ions were captured, cooled and bunched by a different RFQ trap and subsequently analysed using a multi-reflection time-of-flight mass spectrometer[44]. A measured mass spectrum is shown in Extended Data Fig. 1. After 1,000 revolutions in the device, a mass resolving power ($R = m/\Delta m$) of 1.7×10^5 was achieved, which allowed the isobaric beam composition to be analysed. The only mass peak detected was identified as the signal of $^{226}Ra^{19}F^+$, confirming the purity of the beam from ISOLDE.

The intensity of RaF^+ molecules depends strongly on the target and ion source temperature. For a target temperature of 1,300 °C, a mean value of 2×10^7 molecules s^{-1} of $^{226}RaF^+$ was measured after the mass separator. Depending on the molecular mass and beam intensity, the transmission efficiency through the RFQ trap varied from 15% to 30%. The ion-beam transmission from the ion trap to the interaction region was measured to be 25(5)%. The charge exchange cell vapour was heated to produce a measured neutralization rate of 30(5)%. Thus, we estimate that on average 5×10^4 neutral ^{226}RaF molecules s^{-1} were delivered to be resonantly excited. From the analysis of the measured spectra it was concluded that the neutral molecules populate the low-lying vibrational states $v = 0, 1, 2, 3, 4$ following a relative population of 0.47:0.29:0.13:0.05:0.03. Resonantly ionized molecules with rates of the order of 10^3 counts s^{-1} at the peak of the $0 \leftarrow 0$ transition were measured at the particle detector. Future production of RaF^+ molecular rates of the order of 10^9–10^{10} molecules s^{-1} is feasible using active proton irradiation[45].

Laser setup

The resonance ionization schemes used for the study of RaF molecules are shown in Fig. 1. Three different laser systems were prepared to cover the scanning range from 12,800 cm^{-1} to 13,800 cm^{-1}: 1) A dye-laser system (Dye1; Spectrolase 4000, Spectron) provided pulses of 100(5) μJ with a linewidth of 10 GHz (0.3 cm^{-1}). 2) A dye laser (Dye2; Cobra, Sirah) with a narrower linewidth of 2.5 GHz (0.09 cm^{-1}) produced pulses of similar energy. The lasers were loaded with either Styryl 8 or DCM dyes to provide wavenumber ranges 12,800–14,000 cm^{-1} and 15,150–16,600 cm^{-1}, respectively. Both dye lasers were pumped by 532-nm pulses at 100 Hz, obtained from two different heads of a twin-head Nd:YAG laser (LPY 60150–100 PIV, Litron). 3) A grating Ti:sapphire laser system with a linewidth of 2 GHz (0.07 cm^{-1}) produced pulses of 20(1) μJ, pumped by 532-nm pulses at 1 kHz from a Nd:YAG laser (LDP-100MQ, LEE Laser). The non-resonant ionization step was obtained by 355-nm pulses of 30 mJ at 100 Hz, produced by the third-harmonic output of a high-power Nd:YAG laser (TRLi 250-100, Litron).

The release of the ion bunch was synchronized with the laser pulses by triggering the flash-lamps and Q-switch of the pulsed lasers with a digital delay pulse generator (Quantum Composers 9528).

The dye-laser wavelengths were measured with a wavelength meter (WS6-600 HighFinesse) and the Ti:sapphire laser wavelengths were measured by a wavelength meter (WSU-2 HighFinesse) calibrated by measuring a reference wavelength provided by a stabilized diode laser (DLC DL PRO 780, Toptica).

Collinear and anti-collinear excitation

For the initial peak searching, a zero-degree mirror at the end of the beam line was used to reflect the laser light anti-collinearly with respect to the travelling direction of the RaF bunch. Thus, each scanning laser covered two different wavenumber regions in the molecular rest frame, owing to the Doppler shift present for the fast RaF molecules. For a molecule travelling at velocity v, the laser wavenumber in the laboratory frame, $\tilde{\nu}_0$, is related to the wavenumber in the molecule rest frame, $\tilde{\nu}$, by the expression $\tilde{\nu} = \frac{1+\beta\cos\theta}{\sqrt{1-\beta^2}}\tilde{\nu}_0$, with $\beta = v/c$ (c, speed of light in vacuum) and where θ is the angle between the direction of the laser beam and the velocity of the molecule. For RaF molecules at 39,998(1) eV ($v \approx 0.18$ m μs^{-1}), a difference of 15.7 cm^{-1} is obtained between the laser pulse sent out collinearly ($\cos\theta = 1$) and anti-collinearly ($\cos\theta = -1$) with respect to the direction of the velocity of the molecule.

Spectroscopic analysis

The peaks in the different spectra were identified by rebinning the spectra using coarse bin sizes with values up to 1 cm^{-1}. Only groups of data points that were consistently observed with a 5-sigma significance above background were considered as candidates for transitions. The vibrational transitions in Fig. 2 show asymmetric line profiles with a maximum located towards higher wavenumbers. The band centres cannot be determined directly from the measured line profiles, and so we used the wavenumber positions of the maxima in our data analysis. Extended Data Table 1 lists the maximum peak positions and estimated uncertainties are given in parentheses. The wavenumber difference, $\Delta\tilde{\nu}$, of vibrational levels in the electronic $^2\Sigma^+$ ground state and in the $^2\Pi_{1/2}$ excited state were derived from combination differences of the recorded ^{226}RaF spectra (see Extended Data Table 1).

In our analysis we used vibrational energy terms $E_v/(hc)$ of a Morse potential according to:

$$E_v/(hc) = \tilde{\omega}_e\left(v+\frac{1}{2}\right) - \frac{\tilde{\omega}_e^2}{4\tilde{\mathcal{D}}_e}\left(v+\frac{1}{2}\right)^2 \quad (1)$$

Energy-level differences

$$(E_{v+1} - E_v)/(hc) = \tilde{\omega}_e - \frac{\tilde{\omega}_e^2}{2\tilde{\mathcal{D}}_e}(v+1) \quad (2)$$

were used to derive the Morse potential parameters $\tilde{\omega}_e$ and $\tilde{\mathcal{D}}_e$ from a least-squares fit analysis. The derived energy-level differences are given in Extended Data Table 1, whereas Extended Data Table 2 contains the molecular parameters from the fit. The harmonic vibration frequencies $\tilde{\omega}_e$ of the $^2\Sigma^+$ and $^2\Pi_{1/2}$ states are almost identical and correspond well to the theoretical predictions with a deviation of less than 5%; see Extended Data Table 2. The same holds for the estimated dissociation energy $\tilde{\mathcal{D}}_e$, which is in better agreement with the values of ref. [6], as therein also the low-energy part of the potentials was used to estimate the dissociation energy.

In the case of the two low-lying $^2\Pi$ fine-structure levels, the observed origins $T_{0,0}$ agree well with the calculated values based on the Relativistic Correlation Consistent – Atomic Natural Orbital (RCC-ANO) basis set. From the energy difference of the fine-structure components the effective spin-orbital coupling parameter A is derived. For the $^2\Pi$ states, the experimental value of 2,068(5) cm^{-1} is in good agreement with the calculated value. The band origins are in reasonable agreement with results from the RCC-ANO basis set calculation, if one attributes the $\Omega = 3/2$ levels, which were computationally found to be of mixed $\Pi_{3/2}$ and $\Delta_{3/2}$ character in this order of energies. A reverse assignment also gives better agreement with experiment. Calculations of the gas-phase bond lengths, dissociation energies and additional properties of RaF molecules have been reported[8,32,35].

Data availability

Examples of vibronic spectra measured for RaF molecules are included as source data with this Article. All other relevant data supporting the findings of these studies are available from the corresponding author upon request.

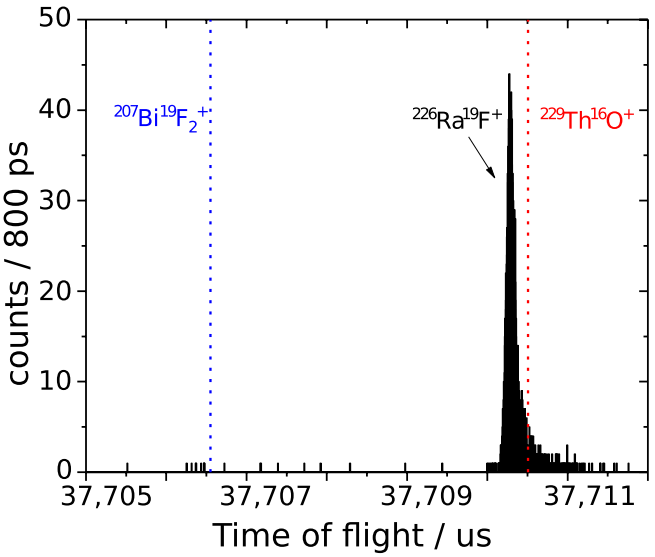
41. Böhlen, T. T. et al. The FLUKA code: developments and challenges for high energy and medical applications. *Nucl. Data Sheets* **120**, 211–214 (2014).
42. *HSC Chemistry 9* v.9.8.1.2 (Outotec, 2018).
43. Kreim, S. et al. Recent exploits of the ISOLTRAP mass spectrometer. *Nucl. Instrum. Meth. B* **317**, 492–500 (2013).
44. Wolf, R. N. et al. Isoltrap's multi-reflection time-of-flight mass separator/spectrometer. *Int. J. Mass Spectrom.* **349-350**, 123–133 (2013).
45. Guglielmetti, A. et al. New measurement of exotic decay of ^{225}Ac by ^{14}C emission. *Eur. Phys. J. A* **12**, 383–386 (2001).

Acknowledgements This work was supported by the ERC Consolidator grant no. 648381 (FNPMLS); Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project no. 328961117 – SFB 1319 ELCH; STFC grants ST/L005794/1, ST/L005786/1 and ST/P004423/1, and Ernest Rutherford grant no. ST/L002868/1; projects from FWO-Vlaanderen, GOA 15/010 from KU Leuven and BriX IAP research programme no. P7/12; European Union grant agreement no. 654002 (ENSAR2); the Russian Science Foundation under grant no. 18-12-00227; and BMBF grants 05P15HGCIA and 05P18HGCIA. We thank J. P. Ramos, J. Ballof and T. Stora for their support in the production of RaF molecules. A.J.B. was suppported by the Henry W. Kendall (1955) Fellowship. We would also like to thank the ISOLDE technical group for their support and assistance. We thank D. Budker for comments and suggestions as well as A. Petrov for discussions on Δ states. R.B. acknowledges I. Tietje for early discussions on various experiments at CERN and thanks A. Welker for sharing knowledge on isotope production and separation as well as for initial discussions on the RaF studies. R.B. acknowledges discussions with K. Gaul on molecular properties and with D. Andrae on finite nuclear size effects. R.B. and T.A.I. acknowledge S. Hoekstra and L. Willmann for early discussions on the production of RaF. T.A.I. thanks A. Zaitsevskii for discussions on the coupled-cluster method.

Author contributions R.F.G.R. led the experiments and R.B. led the theoretical support for this work. R.F.G.R., R.B., C.L.B., M.L.B., K.C., B.S.C., K.T.F., R.P.d.G., S.F., F.P.G., Á.K., H.A.P., C.M.R., S.R., A.R.V., F.W. and S.G.W. performed the experiment. R.F.G.R., R.B., A.A.B., A.J.B. and T.F.G. performed the data analysis. R.F.G.R. prepared the figures. R.B. and T.A.I. performed theoretical predictions that motivated the experimental proposal and analysis of the results. R.F.G.R. and R.B. prepared the initial draft of the manuscript with input from A.A.B., A.J.B., K.T.F., T.F.G., T.A.I., G.N. and S.G.W. All authors discussed the results and contributed to the manuscript at different stages.

Competing interests The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to R.G.R. or R.B.
Peer review information *Nature* thanks Michael Patzschke and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Time-of-flight spectrum measured at mass $A = 245$. The time-of-flight spectrum of the $^{226}RaF^+$ ($A = 245$) beam as delivered from ISOLDE after 1,000 revolutions in the multi-reflection time-of-flight mass spectrometer. A mass resolving power of 1.7×10^5 was achieved, which allowed the isobaric beam composition to be analysed. Only $^{226}Ra^{19}F^+$ ions were detected. The positions of the most probable accompanying ions are highlighted by dotted vertical lines.

Extended Data Table 1 | ^{226}RaF vibrational transitions in the electronic $X^2\Sigma^+$ ground state and the $A^2\Pi_{1/2}$ excited state derived from combination differences

$v' \leftarrow v''$	$^2\Sigma^+$ $\Delta\tilde{\nu}/\mathrm{cm}^{-1}$	$^2\Pi_{1/2}$ $\Delta\tilde{\nu}/\mathrm{cm}^{-1}$
1 - 0	438.4(7)	432.2(7)
2 - 1	435.0(7)	428.9(7)
2 - 1	435.4(11)	429.3(11)
3 - 2	431.6(11)	425.6(7)
4 - 3	428.4(14)	422.2(14)
5 - 4	···	419.2(14)

Combined statistical and systematic uncertainties are given in parentheses.

Extended Data Table 2 | Molecular parameters of RaF from vibrational analysis of the electronic ground state ($X^2\Sigma^+$) and excited states ($A^2\Pi$, $B^2\Delta$ and $C^2\Sigma^+$)

	$\tilde{\omega}_e\,/\mathrm{cm}^{-1}$	$\tilde{T}_e\,/10^4\,\mathrm{cm}^{-1}$	$A\,/10^3\,\mathrm{cm}^{-1}$	$\tilde{\mathcal{D}}_e\,/10^4\,\mathrm{cm}^{-1}$	Ref
$X\,^2\Sigma^+$	441.8(1)			2.92(5)	this work
	432			3.21	[43], theo.[a]
	431			4.26	[43], theo.[b]
$A\,^2\Pi_{1/2}/^2\Pi_{3/2}$	435.5(1)/419.1(2)	1.32878(1)/1.53554(3)	2.0676(36)	2.90(3)/-	this work
	428/410	1.40/1.60	2.0	3.13/-	[43], theo.[a]
	428/415	1.33/1.50	1.7		[43], theo.[b]
$B\,^2\Delta_{3/2}/^2\Delta_{5/2}$	431.9(2)/-	1.51477(2)/-	-/-	2.83(11)/-	this work
	432/419	1.64/1.71	0.4		[43], theo.[a]
	431/423	1.54/1.58	0.2		[43], theo.[b]
$C\,^2\Sigma$	430.9(2)	1.61806(1)		2.78(9)	this work

[a]These calculations used Fock space coupled cluster singles and doubles, the Dyall basis set and a smaller active space.

[b]These calculations used Fock space coupled cluster singles and doubles, the RCC-ANO basis set and a larger active space.

Experimental results are compared with theoretical predictions[43], theo.

For a direct comparison, theoretical values given for $\tilde{\omega}_e$ should be scaled by $\sqrt{m_p/u} \approx 1.0036$ to account for the atomic mass constant instead of the proton mass, m_p, in atomic mass units (u).

Article

Stable, high-performance sodium-based plasmonic devices in the near infrared

https://doi.org/10.1038/s41586-020-2306-9

Received: 3 September 2019

Accepted: 12 March 2020

Published online: 27 May 2020

Check for updates

Yang Wang[1,7], Jianyu Yu[1,7], Yi-Fei Mao[2,7], Ji Chen[1,3,7], Suo Wang[2], Hua-Zhou Chen[2], Yi Zhang[4], Si-Yi Wang[2], Xinjie Chen[1], Tao Li[1,3], Lin Zhou[1,3]✉, Ren-Min Ma[2,5]✉, Shining Zhu[1]✉, Wenshan Cai[6] & Jia Zhu[1]✉

Plasmonics enables the manipulation of light beyond the optical diffraction limit[1–4] and may therefore confer advantages in applications such as photonic devices[5–7], optical cloaking[8,9], biochemical sensing[10,11] and super-resolution imaging[12,13]. However, the essential field-confinement capability of plasmonic devices is always accompanied by a parasitic Ohmic loss, which severely reduces their performance. Therefore, plasmonic materials (those with collective oscillations of electrons) with a lower loss than noble metals have long been sought[14–16]. Here we present stable sodium-based plasmonic devices with state-of-the-art performance at near-infrared wavelengths. We fabricated high-quality sodium films with electron relaxation times as long as 0.42 picoseconds using a thermo-assisted spin-coating process. A direct-waveguide experiment shows that the propagation length of surface plasmon polaritons supported at the sodium–quartz interface can reach 200 micrometres at near-infrared wavelengths. We further demonstrate a room-temperature sodium-based plasmonic nanolaser with a lasing threshold of 140 kilowatts per square centimetre, lower than values previously reported for plasmonic nanolasers at near-infrared wavelengths. These sodium-based plasmonic devices show stable performance under ambient conditions over a period of several months after packaging with epoxy. These results indicate that the performance of plasmonic devices can be greatly improved beyond that of devices using noble metals, with implications for applications in plasmonics, nanophotonics and metamaterials.

Of the plasmonic materials, the noble metals, particularly silver and gold, are those most often used owing to their relatively low loss. However, the optical loss of the two metals is still not commercially acceptable and has been the primary limiting factor for the widespread applications of plasmonics[1–13]. Therefore, there has been a persistent search for low-loss alternatives[14–16], such as crystalline metals, intermetallic composites, metal alloys, nitrides and oxides.

Of these alternatives, the alkali metals, such as sodium, have long been regarded as ideal plasmonic materials[14–17], primarily because of their low intraband damping rate. When light interacts with a plasmonic metal, it suffers from scattering damping of the intraband transition of electrons (γ_p)[15], and the total intraband damping rate can be obtained by:

$$\gamma_p = \gamma_{e\text{-}p} + \gamma_{e\text{-}e} + \gamma_{e\text{-}i}, \quad (1)$$

where $\gamma_{e\text{-}p}$, $\gamma_{e\text{-}e}$ and $\gamma_{e\text{-}i}$ are the optical damping rates that originate from electron–phonon scattering, electron–electron scattering and electron–impurity scattering, respectively. The overall intraband optical loss γ_p of sodium is estimated to be around 0.010 eV (Supplementary Information), corresponding to a relaxation time of 0.42 ps. For comparison, silver has an intraband optical loss of 0.021 eV and a relaxation time of 0.20 ps, based on the Drude–Lorentz model[18]. In addition, sodium has electron gases with a density of 2.65×10^{22} cm^{-3} (ref. [17]), which is approximately half that of silver, which is another important factor contributing to a decreased optical loss.

Although sodium has been predicted to be an ideal plasmonic material for years, the experimental exploration of sodium as a plasmonic material has been limited, apart from the early measurement of its optical constants[19,20] and the demonstration of localized plasmon resonances in alkali metal nanoparticles precipitated within crystal matrices[21]. Because of its high chemical reactivity, fabricating sodium-based structures using conventional metal-deposition techniques, such as physical deposition, has been challenging[19].

Here, by taking advantage of the low melting point of sodium, we have developed a thermo-assisted spin-coating process for fabricating a sodium film, as shown in Fig. 1a (optical images of the procedure in Extended Data Fig. 1a). The sodium metal is heated up to 160 °C to

[1]National Laboratory of Solid State Microstructures, Collaborative Innovation Center of Advanced Microstructures, Jiangsu Key Laboratory of Artificial Functional Materials, College of Engineering and Applied Sciences, Nanjing University, Nanjing, People's Republic of China. [2]State Key Lab for Mesoscopic Physics and School of Physics, Peking University, Beijing, People's Republic of China. [3]Key Laboratory of Intelligent Optical Sensing and Manipulation, Ministry of Education, Nanjing University, Nanjing, People's Republic of China. [4]School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou, People's Republic of China. [5]Frontier Science Center for Nano-optoelectronics and Collaborative Innovation Center of Quantum Matter, Peking University, Beijing, People's Republic of China. [6]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [7]These authors contributed equally: Yang Wang, Jianyu Yu, Yi-Fei Mao, Ji Chen. ✉e-mail: linzhou@nju.edu.cn; renminma@pku.edu.cn; zhusn@nju.edu.cn; jiazhu@nju.edu.cn

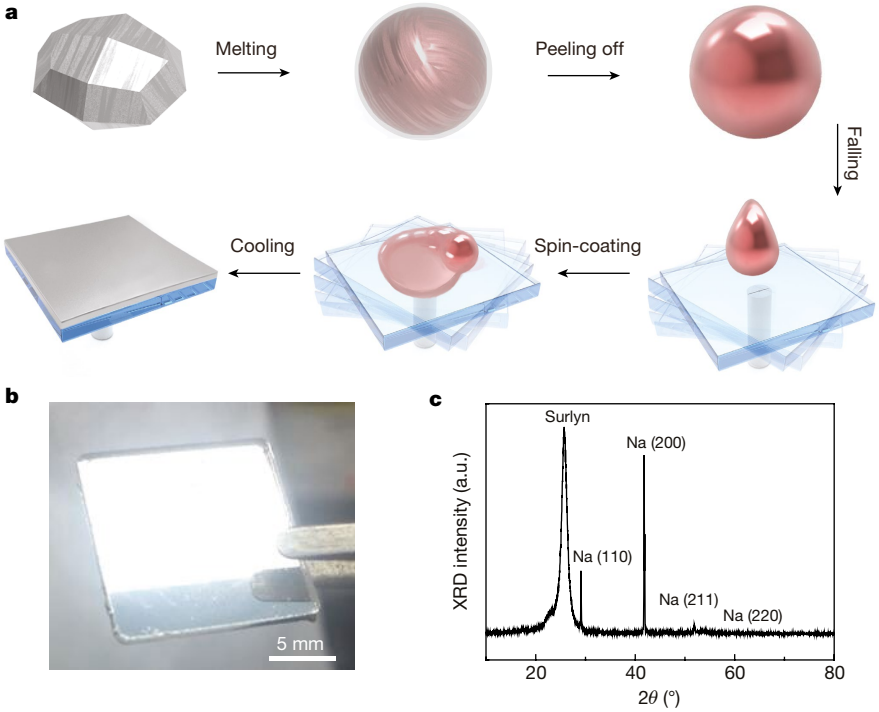


Fig. 1 | Sodium film fabricated by a thermo-assisted spin-coating process. **a**, Schematic of the process flow. A piece of sodium is heated and melted while inner impurities diffuse towards the surface during this thermal annealing process; the thin oxide shell incorporating the impurities covers the melted sodium droplet and is then peeled off; the cleaned sodium droplet is spin-coated onto a rotating quartz substrate; the sodium film then forms on the quartz surface with a smooth interface. **b**, Photograph of the spin-coated sodium film on quartz. Scale bar, 5 mm. **c**, X-ray diffraction pattern of the spin-coated sodium film. The sodium film is covered with Surlyn as packaging. XRD, X-ray diffraction; a.u., arbitrary units.

form a droplet. During this thermal annealing process, impurities diffuse towards the surface, which can then be peeled off to purify the sodium droplet. The purified sodium droplet, which has a luminous appearance, is then spin-coated onto an ultrasmooth quartz substrate (roughness of about 0.1 nm; Extended Data Fig. 1b). Once the sodium droplet touches the surface of spinning quartz, an ultrasmooth sodium film forms (Fig. 1b) owing to the strong centrifugal forces of the rotating quartz and fast solidification. The entire process is conducted inside a glove box with an inert atmosphere. The X-ray diffraction pattern shows that the sodium film produced is polycrystalline (Fig. 1c, where the sample is sealed with Surlyn, an ionomer resin transparent to X-rays for a wide range of angles).

To evaluate the optical properties of the prepared sodium film, we measured its dielectric function in the wavelength range 400–1,500 nm using a spectroscopic ellipsometer. Figure 2a depicts the measured dielectric functions versus wavelength of the sodium film (solid circle symbols), which exhibits a lower optical loss than reported values for silver over a wide range of wavelengths (see Extended Data Fig. 2 for comparison).

To quantitatively analyse the loss mechanisms of the sodium film, we used a Drude–Lorentz model[16,22] to fit the measured dielectric curves (dashed lines in Fig. 2a), which is expressed as:

$$\varepsilon(\omega) = \varepsilon_b - \omega_p^2/(\omega^2 + \mathrm{i}\omega\gamma_p) + f_1\omega_1^2/(\omega_1^2 - \omega^2 - \mathrm{i}\omega\gamma_1), \quad (2)$$

where ε_b is the polarization response from the core electrons (background permittivity), ω_p is the bulk plasma frequency, γ_p is the Drude damping rate, f_1 and ω_1 are the amplitude and resonant frequency of the inter-band transition, respectively, and γ_1 is the related interband damping rate. The fitting parameters for the sodium film are $\varepsilon_b = 0.500$, $\omega_p = 5.414$ eV, $\omega_1 = 2.945$ eV, $f_1 = 0.280$, $\gamma_1 = 2.706$ eV and $\gamma_p = 0.010$ eV. The relaxation time τ can be determined by the Drude damping rate, which is 0.42 ps. Notably, compared to the bulk silver materials, the optical damping rate of our sodium film is reduced by half and the relaxation time is doubled. ($\gamma_p = 0.021$ eV and $\tau = 0.20$ ps for silver[18]).

To further illustrate the reduced optical loss of our sodium film compared to silver, we calculated the figure of merit (defined by $-\varepsilon_1/\varepsilon_2$; ref. [16]), as shown in Fig. 2b. We chose two sets of silver data from recent publications for comparison[23,24]. It can clearly be seen that our sodium film has state-of-the-art performance in the near-infrared region of the spectrum.

This low-loss sodium film provides an opportunity to improve the performance of plasmonic devices. We now demonstrate high-performance sodium-based plasmonic waveguides and plasmonic nanolasers at near-infrared wavelengths. A two-step fabrication process—spin-coating of sodium and packaging—was conducted inside a glove box. All characterizations were carried out under ambient conditions, including the measurements of the dielectric function, plasmonic waveguiding and plasmonic lasing.

Figure 2c shows the schematics of the sodium-based plasmonic waveguide structure. The left coupler converts the incident laser beam to surface plasmon polaritons (SPPs) (Fig. 2d, left light spot), which propagate along the sodium–quartz interface and then become coupled to free space via the right coupler (Fig. 2d, right light spot). The intensity of the SPPs decreases exponentially as a function of distance. Various propagation separations were measured to fit the intensity decay curve (Fig. 2e), to obtain the propagation lengths (δ_{SPP}) at different wavelengths on the sodium–quartz interface (Fig. 2f). The inset to Fig. 2e depicts the scanning electron microscopy (SEM) image of the launching and outcoupling structures for the SPP propagation measurements (the fabrication process for the plasmonic waveguide devices is described in Extended Data Fig. 3). It is clear that the sodium film supports a δ_{SPP} of 200 μm at a wavelength of 1,500 nm, which represents state-of-the-art performance for plasmonic waveguides (see Extended Data Fig. 4a–c for comparison with silver).

In addition to the propagation length, the effective mode size—expressed as the sum of the decay lengths (skin depths) in the dielectric

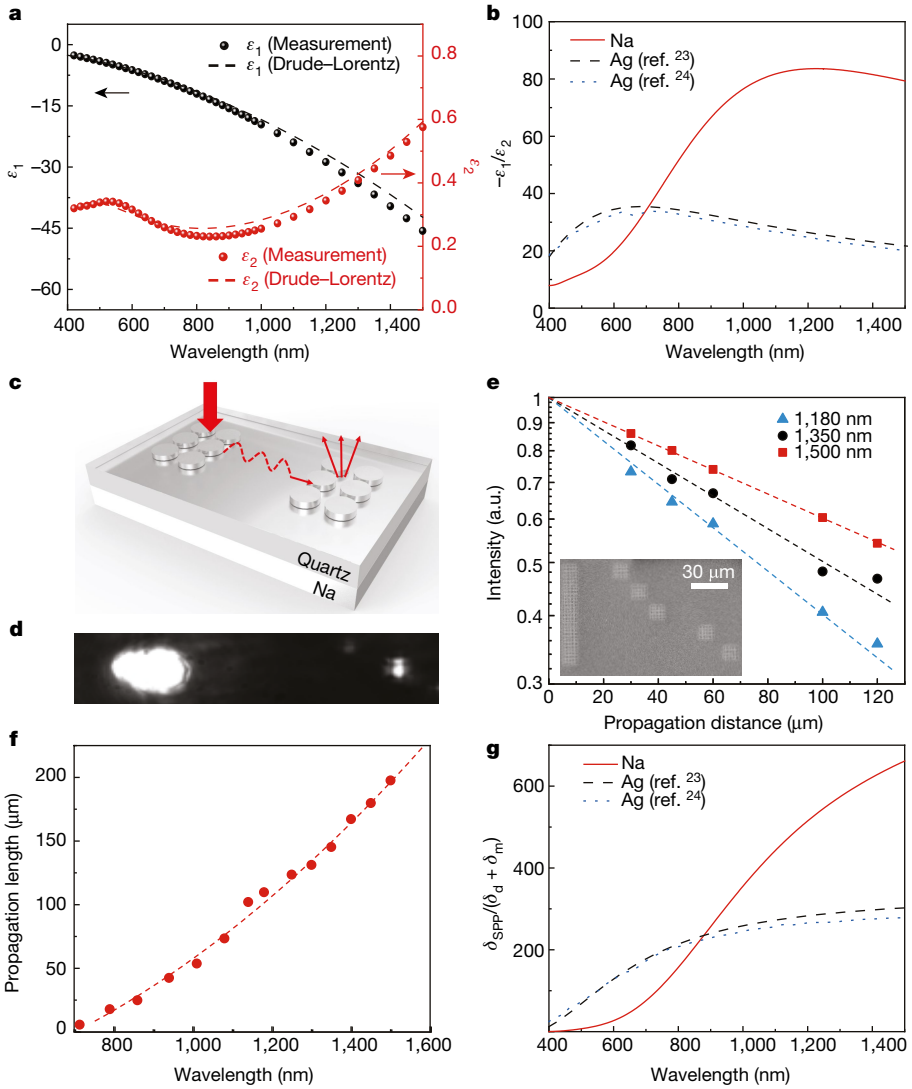


Fig. 2 | Dielectric functions of sodium film and sodium-based plasmonic waveguides. a, Real and imaginary parts of the dielectric functions ($\varepsilon = \varepsilon_1 + i\varepsilon_2$) of the sodium film (solid circles). Dashed lines show the Drude–Lorentz fittings. **b**, Figure of merit ($-\varepsilon_1/\varepsilon_2$) comparison for sodium (this work) and silver(from refs. [23,24]). Comparisons with various data for silver in the literature are shown in Extended Data Fig. 4. **c**, Schematic of a sodium-based plasmonic waveguiding structure with two nanostructured couplers. The coupling structures are pillar arrays. **d**, Optical image of the light spots of sodium-based plasmonic waveguiding. The distance between the launching and output structures is 100 μm. **e**, Propagation measurements at wavelengths of 1,180 nm, 1,350 nm and 1,500 nm for sodium, with exponential curves fitted to the data. Inset, SEM image of the launching and outcoupling structures without sodium coating. Scale bar, 30 μm. **f**, The fitted propagation length at different wavelengths on the sodium–quartz interface. The dashed line is a fit to the data. **g**, Figure of merit for plasmonic waveguides for sodium and silver, which is defined as the ratio of propagation length to skin depth $\delta_{SPP}/(\delta_d + \delta_m)$.

δ_d and in the metal δ_m—is another essential parameter of the plasmonic waveguide, characterizing the field-confinement capability. Therefore, we can define the figure of merit of the plasmonic waveguide as the ratio of the propagation length over skin depth $\delta_{SPP}/(\delta_d + \delta_m)$. The figures of merit of the plasmonic waveguide of sodium and silver are plotted in Fig. 2g, further identifying sodium as a promising material of choice for plasmonic-waveguide-based applications at the near-infrared wavelengths (see Extended Data Fig. 4d for comparison with more data for silver).

The plasmonic nanolaser is another widely studied plasmonic device for which a low-loss metal has long been sought in order to achieve a lower lasing threshold and power consumption[7,25–30]. We fabricated sodium-based plasmonic nanolasers based on a metal–insulator–semiconductor gap plasmonic mode configuration[31,32]. As shown in Fig. 3a, b, the device consists of an InGaAsP multi-quantum-wells (MQWs) nanodisk on top of sodium film with a 7-nm-thick Al_2O_3 layer in between (see Extended Data Fig. 5 for fabrication details). Figure 3c shows the SEM image of an InGaAsP MQWs nanodisk without the sodium coating, with diameter and thickness 1.2 μm and 200 nm, respectively. Figure 3d–f shows the simulated electric-field distributions of the lasing plasmonic mode obtained by three-dimensional full-wave simulations. The electric field is strongly confined at the interface between InGaAsP and sodium, thus exhibiting a pronounced plasmonic feature.

The sodium-based plasmonic nanolaser exhibits single-mode lasing with a low lasing threshold under optical pumping (Fig. 3g). Notably, pronounced resonance peaks appear in the spontaneous emission spectrum below the lasing threshold, which indicates the high quality factor Q of the sodium-based plasmonic nanocavity[31,32]. When we pump the device above the lasing threshold, a single lasing mode becomes dominant at 1,257 nm with a much narrower linewidth than spontaneous emission. The side-mode suppression ratio of this laser is about 20 dB (pump intensity 475 kW cm^{-2}; see Extended Data Fig. 6). We have identified this lasing cavity mode as the gap plasmonic whispering-gallery mode with an azimuthal order of $l = 15$ and a cold

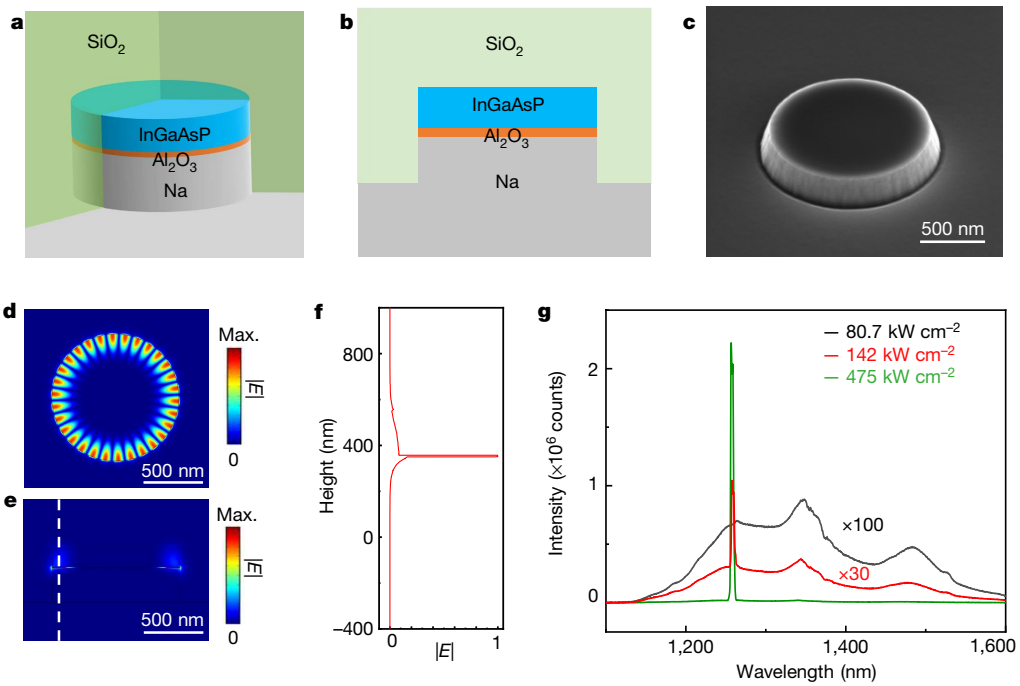


Fig. 3 | Room-temperature sodium-based plasmonic nanolaser. **a**, **b**, Schematics of a sodium-based plasmonic nanolaser in three dimensions (**a**) and two dimensions (**b**). **c**, SEM image of an InGaAsP MQWs nanodisk, without sodium coating. Scale bar, 500 nm. **d**, **e**, Top (**d**) and side (**e**) views of the electric-field distribution of the calculated lasing mode. The mode is a plasmonic whispering-galley mode where the field is strongly confined at the interface between the MQWs and sodium. **f**, Normalized electric-field profile along the white dashed line in **e**. **g**, Spectra of spontaneous emission (black), amplified spontaneous emission (red), and single-mode laser emission (green) of the sodium-based plasmonic nanolaser.

cavity quality factor of about 340 using full-wave simulation (see Extended Data Fig. 7).

The threshold of the laser can be extracted from the evolution of the normalized spectra of the device versus the pump power and the S-shaped light–light curve (Fig. 4a, b). The threshold of the sodium-based plasmonic nanolaser is about 140 kW cm^{-2}, which, to our knowledge, is the lowest reported value among near-infrared plasmonic nanolasers at room temperature (see Extended Data Table 1 and Extended Data Fig. 8).

Owing to the high material dispersion, plasmonic modes possess a higher group index than photonic modes. Here we reveal photonic-to-plasmonic mode jumping by tracking the group index of the lasing modes with decreased cavity diameter (Fig. 4c). In the experiment, we observe a sudden clear increase of the group index at a device diameter of about 2 μm, which indicates that plasmonic mode lasing becomes dominant for smaller cavities. To further demonstrate the effect of sodium in lasing, we carried out a control experiment where bare InGaAsP MQWs nanodisks are optically pumped. However, for the devices with the same diameter as the one shown in Fig. 3, no lasing behaviour is observed till they are burned by the pump laser, which results from the larger radiation loss for the bare dielectric cavities[33].

One natural concern with regard to the use of sodium for plasmonic devices is its stability. It is therefore encouraging to find that these sodium-based plasmonic devices remain stable over a long period of time after the packaging of quartz and epoxy. More details of our extensive accelerated-ageing tests of sodium-based devices (high temperature and high air humidity) can be found in Extended Data Fig. 9. Our nanolaser devices remain functional at a low threshold even after six months (Extended Data Fig. 9g).

In summary, we have demonstrated a method of fabricating high-quality sodium films. Sodium-based plasmonic waveguides and nanolasers have been realized and their high performance is verified

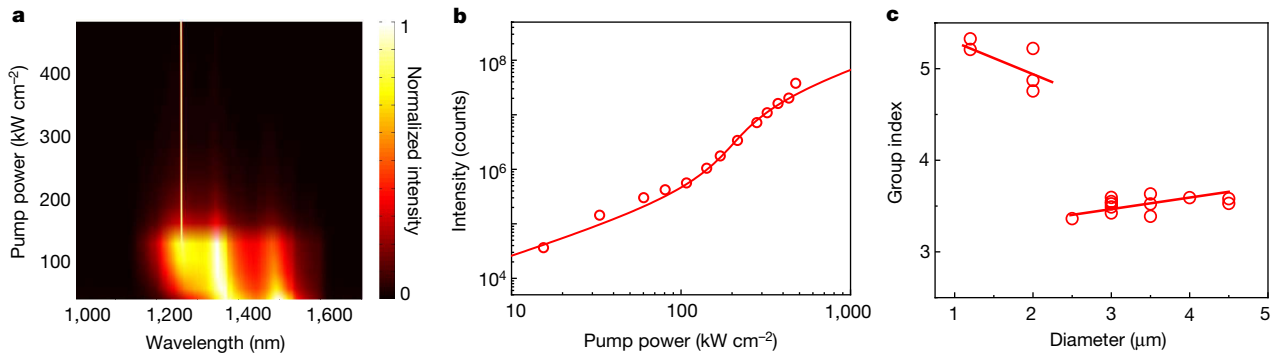


Fig. 4 | Lasing characteristics of sodium-based plasmonic nanolasers. **a**, Normalized spectra of the sodium-based plasmonic nanolaser at different pump powers. **b**, Light–light curve of the plasmonic nanolaser. The circle symbols represent experimental data and the line is a fit to the data. The substantially narrowed emission linewidth above the lasing threshold in **a** and the S-shaped light–light curve in **b** show a clear phase transition from spontaneous emission to lasing emission of the device. **c**, The group index versus device diameter estimated from the mode spacing in lasing spectra.

by systematic comparison with the reported devices based on noble metals. Furthermore, we show that these sodium-based plasmonic devices can operate stably over several months. Our results provide an alternative pathway to low-loss plasmonic materials. The convenient fabrication process and the state-of-the-art performance of the sodium-based plasmonic devices open up opportunities for advanced plasmonic applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2306-9.

1. Maier, S. A. *Plasmonics: Fundamentals and Applications* (Springer Science and Business Media, 2007).
2. Ozbay, E. Plasmonics: merging photonics and electronics at nanoscale dimensions. *Science* **311**, 189–193 (2006).
3. Gramotnev, D. K. & Bozhevolnyi, S. I. Plasmonics beyond the diffraction limit. *Nat. Photon.* **4**, 83–91 (2010).
4. Schuller, J. A. et al. Plasmonics for extreme light concentration and manipulation. *Nat. Mater.* **9**, 193–204 (2010).
5. Fang, Y. R. & Sun, M. T. Nanoplasmonic waveguides: towards applications in integrated nanophotonic circuits. *Light Sci. Appl.* **4**, 294 (2015).
6. Atwater, H. A. & Polman, A. Plasmonics for improved photovoltaic devices. *Nat. Mater.* **9**, 205–213 (2010).
7. Ma, R. M. & Oulton, R. F. Application of nanolasers. *Nat. Nanotechnol.* **14**, 12–22 (2019).
8. Valentine, J. et al. Three-dimensional optical metamaterial with a negative refractive index. *Nature* **455**, 376–379 (2008).
9. Gabrielli, L. H. et al. Silicon nanostructure cloak operating at optical frequencies. *Nat. Photon.* **3**, 461–463 (2009).
10. Lal, S., Link, S. & Halas, N. J. Nano-optics from sensing to waveguiding. *Nat. Photon.* **1**, 641–648 (2007).
11. Chung, T., Lee, S. Y., Song, E. Y., Chun, H. & Lee, B. Plasmonic nanostructures for nano-scale bio-sensing. *Sensors* **11**, 10907–10929 (2011).
12. Fang, N., Lee, H., Sun, C. & Zhang, X. Sub-diffraction-limited optical imaging with a silver superlens. *Science* **308**, 534–537 (2005).
13. Liu, Z. W., Lee, H., Xiong, Y., Sun, C. & Zhang, X. Far-field optica hyperlens magnifying sub-diffraction-limited object. *Science* **315**, 1686 (2007).
14. Boltasseva, A. & Atwater, H. A. Low-loss plasmonic metamaterials. *Science* **331**, 290–291 (2011).
15. Khurgin, J. B. How to deal with the loss in plasmonics and metamaterials. *Nat. Nanotechnol.* **10**, 2–6 (2015).
16. Naik, G. V., Shalaev, V. M. & Boltasseva, A. Alternative plasmonic materials: beyond gold and silver. *Adv. Mater.* **25**, 3264–3294 (2013).
17. Palik, E. D. *Handbook Of Optical Constants Of Solids* Vol. 3 (Academic Press, 1998).
18. Johnson, P. B. & Christy, R. W. Optical constants of the noble metals. *Phys. Rev. B* **6**, 4370–4379 (1972).
19. Smith, N. V. Optical constants of sodium and potassium from 0.5 to 4.0 eV by split-beam ellipsometry. *Phys. Rev.* **183**, 634–644 (1969).
20. Inagaki, T., Emerson, L. C., Arakawa, E. T. & Williams, M. W. Optical properties of solid Na and Li between 0.6 and 3.8 eV. *Phys. Rev. B* **13**, 2305–2313 (1976).
21. Ge, Y. C. Optical absorption study on the annealing behavior of metallic colloids in KCl:K crystals. *Spectrosc. Lett.* **31**, 123–134 (1998).
22. Chen, W., Thoreson, M. D., Ishii, S., Kildishev, A. V. & Shalaev, V. M. Ultra-thin ultra-smooth and low-loss silver films on a germanium wetting layer. *Opt. Express* **18**, 5124–5134 (2010).
23. Yang, H. U. et al. Optical dielectric function of silver. *Phys. Rev. B* **91**, 235137 (2015).
24. Mori, T. et al. Optical properties of low-loss Ag films and nanostructures on transparent substrates. *ACS Appl. Mater. Interf.* **10**, 8333–8340 (2018).
25. Ma, R. M. Lasing under ultralow pumping. *Nat. Mater.* **18**, 1152–1153 (2019).
26. Ning, C. Z. Semiconductor nanolasers and the size-energy-efficiency challenge: a review. *Adv. Photon.* **1**, 014002 (2019).
27. Wang, D., Wang, W., Knudson, M. P., Schatz, G. C. & Odom, T. W. Structural engineering in plasmon nanolasers. *Chem. Rev.* **118**, 2865–2881 (2018).
28. Gwo, S. & Shih, C. K. Semiconductor plasmonic nanolasers: current status and perspectives. *Rep. Prog. Phys.* **79**, 086501 (2016).
29. Hill, M. T. & Gather, M. C. Advanced in small lasers. *Nat. Photon.* **8**, 908–918 (2014).
30. Berini, P. & De Leon, I. Surface plasmon-polariton amplifiers and lasers. *Nat. Photon.* **6**, 16–24 (2012).
31. Ma, R. M., Oulton, R. F., Sorger, V. J., Bartal, G. & Zhang, X. Room temperature sub-diffraction-limited plasmon laser by total internal refection. *Nat. Mater.* **10**, 110–113 (2011).
32. Wang, S., Chen, H. Z. & Ma, R. M. High performance plasmonic nanolasers with external quantum efficiency exceed 10%. *Nano Lett.* **18**, 7942–7948 (2018).
33. Wang, S. et al. Unusual scaling laws for plasmonic nanolasers beyond the diffraction limit. *Nat. Commun.* **8**, 1889 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Thermo-assisted spin-coating process of low-loss sodium film

Extended Data Fig. 1a shows the thermo-assisted spin-coating procedure for the sodium film. The sodium brick was melted to form a droplet on a tungsten boat when it was heated up to 160 °C in an inert-atmosphere-equipped glove box. The oxide shell on the surface of the droplet can be peeled off using tweezers. The liquid sodium can then be dropped onto a fast-spinning quartz substrate on a spin coater. The quartz substrate is ultrasmooth (root-mean-square surface roughness of about 0.1 nm), as shown in Extended Data Fig. 1b. The sodium–glass interface was then prepared. The sodium film, once packaged with epoxy, can be transferred freely out of the glove box.

Dielectric constant characterization

The dielectric constant ε of the sodium film was measured using an ellipsometer (RC2 UI, J. A. Woollam) ranging from 400 nm to 1,500 nm (see Extended Data Fig. 2a). The incident light through the silica substrate can be separated into two light spots; the detector receives only the light from the metal–silica interface. We use the Lorentz model (equations (3)–(5)) to fit the measured data. For the ultraviolet pole and the infrared pole, ε is expressed as

$$\varepsilon_{\mathrm{pole},n} = C_n/(E_n^2 - E^2), \tag{3}$$

where C_n (in units of eV^2) and E_n (in units of eV) are fitting parameters, E is the photon energy (in units of eV) and n is the oscillator number. For the general Lorentz oscillator, ε is expressed as

$$\varepsilon_{\mathrm{Lorentz},m} = A_m B_m E_m/(E_n^2 - E^2 - \mathrm{i}EB_m). \tag{4}$$

Here A_m (unitless), B_m (in units of eV) and E_m (in units of eV) are the fitting parameters, where A_m approximately equals ε_2 at its peak value and B_m is approximately the full-width at half-maximum. m is the oscillator number. In the full wavelength range, the overall dielectric function is

$$\varepsilon(E) = \varepsilon_1(E) - \mathrm{i}\varepsilon_2(E) = \varepsilon_{\mathrm{offset}} + \varepsilon_{\mathrm{pole},n} + \varepsilon_{\mathrm{Lorentz},m}, \tag{5}$$

where $\varepsilon_{\mathrm{offset}}$ and $\varepsilon_{\mathrm{pole},n}$ are real, and contribute only to ε_1. The oscillator functions are complex and therefore contribute to both ε_1 and ε_2.

Comparison of real and imaginary parts of dielectric functions of sodium and silver

Extended Data Fig. 2b, c shows a comparison of the dielectric functions of sodium and silver. Clearly, the optical loss (the imaginary part) of sodium is lower than that for silver, especially in the infrared region. We further compare the figure of merit ($-\varepsilon_1/\varepsilon_2$) in Extended Data Fig. 2d. The data for silver are measurements taken from the literature[17,18,23,24,34–37].

Fabrication of nanostructured sodium film

A thin Ag film (about 30 nm) was evaporated onto the quartz (thickness about 0.2 mm) as a conductive layer using physical vapour deposition (Gatan 682), and double hole arrays were made on the quartz substrate via focused-ion-beam milling (Dual-beam FIB 235, FEI Strata). After the focused-ion-beam process, the Ag film on the surface of the quartz was removed with HNO_3, followed by the thermo-assisted spin-coating process for the sodium film described above.

Measurement of propagation of SPP mode on sodium-quartz interface

The fabrication process for the plasmonic waveguide devices is shown in Extended Data Fig. 3. For effective coupling and decoupling of SPP along the sodium–quartz interface, the quartz substrate was first milled with a focused ion beam to generate periodic patterns at both coupling and out-coupling positions. The laser (Fianium SC-400-4 Compact) was concentrated onto the sodium nanostructures by an objective lens. Coupled-out signals was also captured by the objective lens. Finally, both input and output signals were collected with an infrared charge coupled device (XEVA-1083, Xenics) by a beam splitter. We measured several different propagation separations and fitted the intensity decay curve to unambiguously obtain the propagation lengths at different wavelengths on the sodium–quartz interface.

Comparison of sodium plasmonic waveguides with silver-based ones

We fabricated and measured the δ_{SPP} of both sodium and silver plasmonic waveguide as shown in Extended Data Fig. 4a–c. The sodium supports longer δ_{SPP} than silver in the wavelength range >1 μm. It should be noted that, besides δ_{SPP}, the effective mode size is another essential parameter of a plasmonic waveguide, characterizing the field-confinement capability. It is represented by the decay lengths (skin depths) in the dielectric δ_{d} and in the metal δ_{m}. Therefore, for evaluating the plasmonic waveguide, comparing δ_{SPP} alone is not comprehensive, and the figure of merit of the plasmonic waveguide ($\delta_{\mathrm{SPP}}/(\delta_{\mathrm{d}} + \delta_{\mathrm{m}})$) is usually used. In Fig. 2g we benchmarked our sodium plasmonic waveguide with the two silver-based references. To further clarify this point, we have calculated the figure of merit of the plasmonic waveguide for Ag plasmonic waveguides from the dielectric functions reported in a number of representative publications[17,18,23,24,34–37]. From Extended Data Fig. 4d, one can unambiguously come into the conclusion that our sodium-based plasmonic waveguide has state-of-the-art performance.

Detailed fabrication process for plasmonic nanolasers

The detailed fabrication process of the plasmonic nanolasers is shown in Extended Data Fig. 5. Electron-beam lithography and inductively coupled plasma etching were used to pattern InGaAsP MQWs nanodisks onto an epi-wafer (Extended Data Fig. 5a, b). 500-nm-thick SiO_2 was deposited on the nanodisks by chemical vapour deposition (Extended Data Fig. 5c). Then the whole sample was bonded onto quartz using a benzocyclobutene adhesive layer (Extended Data Fig. 5d), after which the InP substrate was removed by wet etching using HCl solution. Atomic layer deposition was used to deposit 7-nm-thick Al_2O_3 onto the nanodisk (Extended Data Fig. 5e). Finally, the sodium film was spin-coated onto nanodisks and the sample was packaged onto quartz by epoxy (Extended Data Fig. 5f).

Optical characterization of plasmonic nanolasers

Sodium-based plasmonic nanolasers were optically pumped by a nanosecond pump laser (1,064 nm, pulse length 5 ns, repetition rate 12 kHz), focused on the sample with an objective lens (50×, numerical aperture 0.42). The pump laser spot is about 7 μm in diameter. The emission from a nanolaser was collected by the same objective and guided to an InGaAs infrared camera and a near-infrared spectrometer. All experiments were carried out at room temperature.

Numerical simulation of the sodium-based plasmonic nanolaser

We carried out three-dimensional full-wave simulations via commercial software COMSOL multiphysics (radio-frequency module) to calculate the optical modes of the nanolasers. The Q factors of the cavity modes were evaluated as $Q = f_{\mathrm{r}}/\delta f$, where f_{r} and δf represent the resonance frequency and the full-width at half-maximum of the resonance spectrum, respectively.

Comparison of sodium-based plasmonic nanolasers with noble-metal-based ones

Extended Data Table 1 shows a comparison of sodium-based plasmonic nanolasers with reported noble-metal-based ones operated in the near-infrared range[38–49]. In the table, the feedback mechanism, physical size, gain medium, emission wavelength, pump condition, working temperature and lasing threshold in peak power density are

listed for comparison. Extended Data Fig. 8 shows the lasing thresholds and physical volumes of reported room-temperature plasmonic nanolasers in the near-infrared range.

Data availability

We declare that the data supporting the findings of this study are available within the paper.

34. McPeak, K. M. et al. Plasmonic films can easily be better: rules and recipes. *ACS Photon*. **2**, 326–333 (2015).
35. Babar, S. & Weaver, J. H. Optical constants of Cu, Ag, and Au revisited. *Appl. Opt*. **54**, 477–481 (2015).
36. Werner, W. S. M., Glantschnig, K. & Ambrosch-Draxl, C. Optical constants and inelastic electron-scattering data for 17 elemental metals. *J. Phys. Chem. Ref. Data* **38**, 1013–1092 (2009).
37. Ciesielski, A., Skowronski, L., Trzinski, M. & Szoplik, T. Controlling the optical parameters of self-assembled silver films with wetting layers and annealing. *Appl. Surf. Sci*. **421B**, 349–356 (2017).
38. Hill, M. T. et al. Lasing in metal–insulator–metal sub-wavelength plasmonic waveguides. *Opt. Express* **17**, 11107–11112 (2009).
39. Khajavikhan, M. et al. Thresholdless nanoscale coaxial lasers. *Nature* **482**, 204–207 (2012).
40. Zhou, W. et al. Lasing action in strongly coupled plasmonic nanocavity arrays. *Nat. Nanotechnol*. **8**, 506–511 (2013).
41. Yang, A. et al. Real-time tunable lasing from plasmonic nanocavity arrays. *Nat. Commun*. **6**, 6939 (2015).
42. Keshmarzi, E. K., Tait, R. N. & Berini, P. Single-mode surface plasmon distributed feedback lasers. *Nanoscale* **10**, 5914–5922 (2018).
43. Kwon, S.-H. et al. Subwavelength plasmonic lasing from a semiconductor nanodisk with silver nanopan cavity. *Nano Lett*. **10**, 3679–3683 (2010).
44. Marell, M. J. H. et al. Plasmonic distributed feedback lasers at telecommunications wavelengths. *Opt. Express* **19**, 15109–15118 (2011).
45. Kwon, S.-H. et al. Surface plasmonic nanodisk/nanopan lasers. *IEEE J. Quantum Electron*. **47**, 1346–1353 (2011).
46. Lakhani, A. M. et al. Plasmonic crystal defect nanolaser. *Opt. Express* **19**, 18237–18245 (2011).
47. Ho, J. et al. Low-threshold near-infrared GaAs-AlGaAs core-shell nanowire plasmon laser. *ACS Photon*. **2**, 165–171 (2015).
48. Ho, J. et al. A nanowire-based plasmonic quantum dot laser. *Nano Lett*. **16**, 2845–2850 (2016).
49. Lee, C.-J. et al. Low-threshold plasmonic lasers on a single-crystalline epitaxial silver platform at telecom wavelength. *ACS Photon*. **4**, 1431–1439 (2017).

Acknowledgements We thank H. Su and James N. Hilfiker at Genuine Optronics for help with ellipsometry measurements. We acknowledge the microfabrication centre of the National Laboratory of Solid State Microstructures (NLSSM) for technical support. This work is jointly supported by the National Key Research and Development Program of China (grant numbers 2017YFA0205700, 2018YFA0704401), National Natural Science Foundation of China (grant numbers 51925204, 11874211, 11621091, 61735008, 11774014, 11574012, 61521004 and 91950115), Beijing Natural Science Foundation (grant number Z180011) and the Fundamental Research Funds for the Central Universities (grant numbers 021314380140 and 021314380150).

Author contributions J.Z., R.-M.M. and S.Z. conceived and supervised the project. Y.W., J.Y., Y.-F.M. and X.C. fabricated the samples. J.Y. collected dielectric function data. Y.W., J.Y., J.C. and T.L. acquired the surface plasma propagation data. S.W., S.-Y.W. and Y.-F.M. collected and analysed lasing data. H.-Z.C. performed the lasing mode simulations. Y.Z. and L.Z. calculated the plasmon damping rate. Y.W., J.Y., S.W., J.Z., R.-M.M. and W.C. wrote the paper with input from all authors.

Competing interests The authors declare no competing interests.

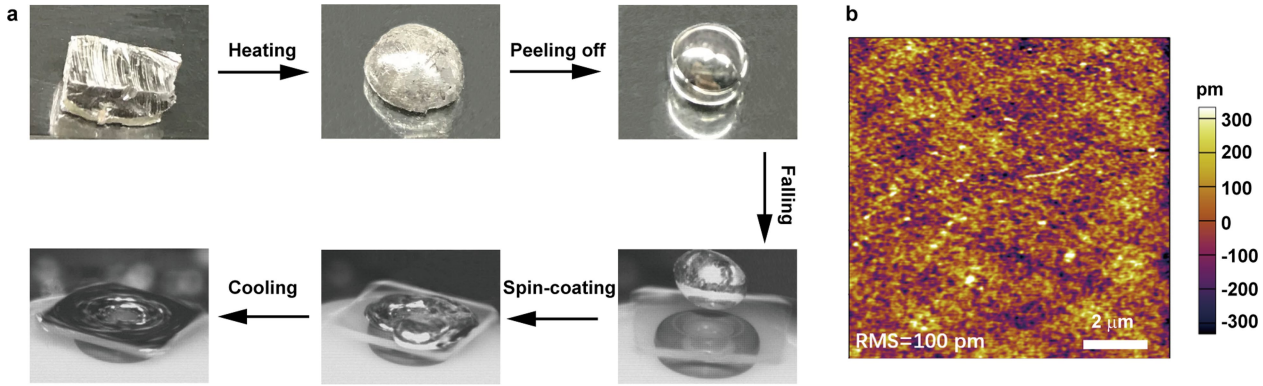
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2306-9.

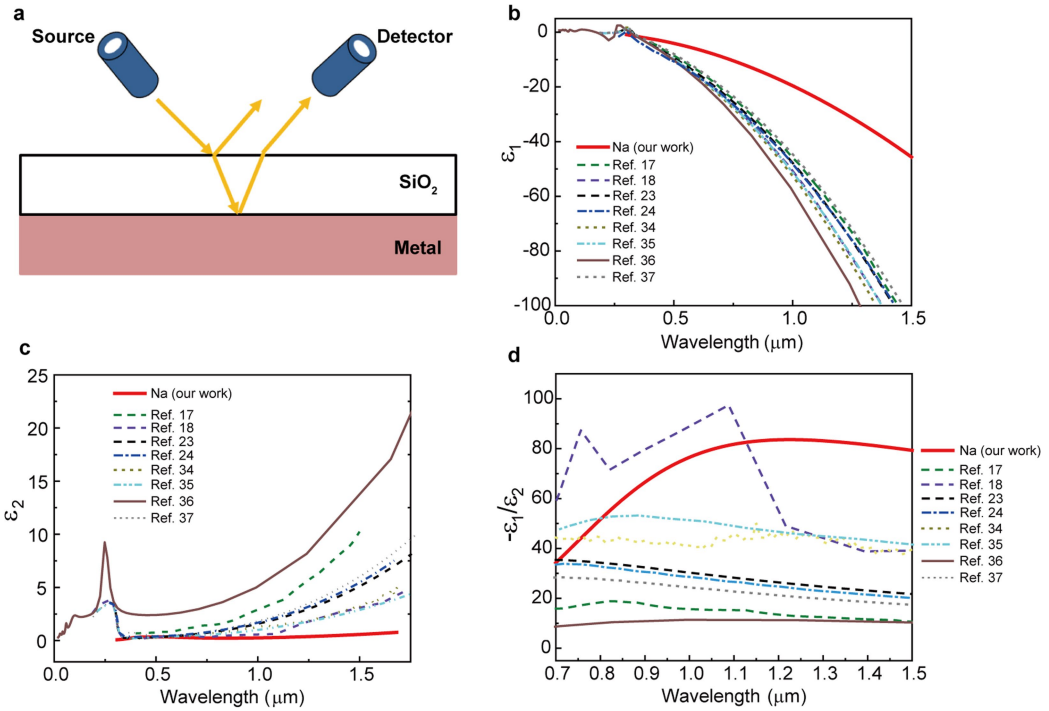
Correspondence and requests for materials should be addressed to L.Z., R.-M.M., S.Z. or J.Z.

Peer review information *Nature* thanks Michael Cortie, Gururaj Naik and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

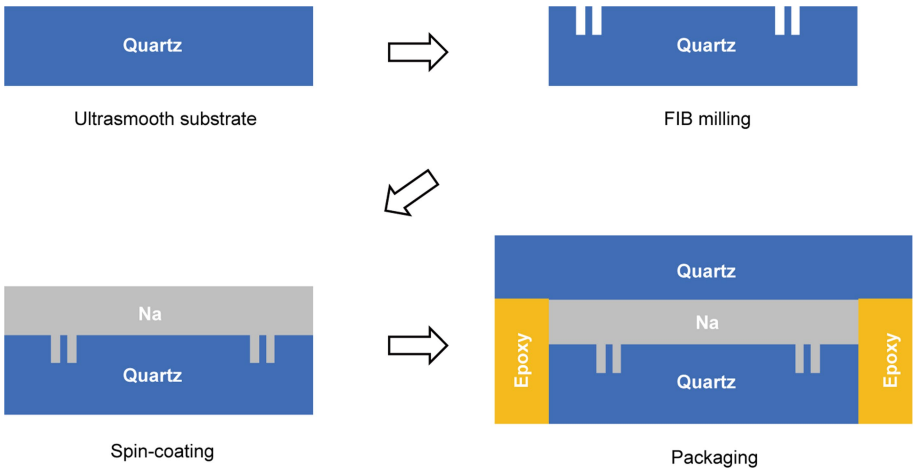
Reprints and permissions information is available at http://www.nature.com/reprints.



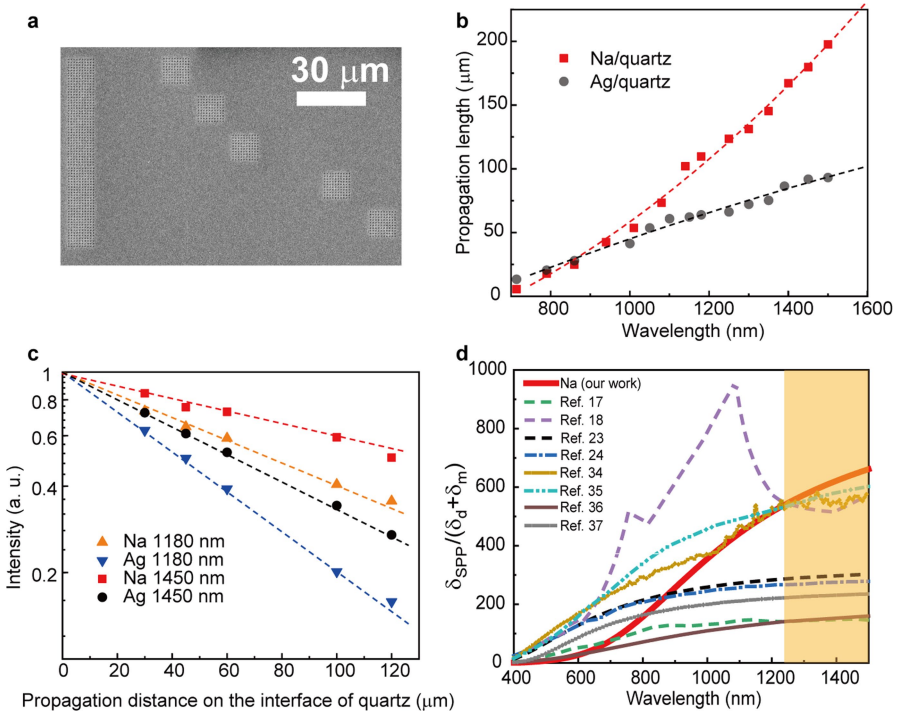
Extended Data Fig. 1 | The thermo-assisted spin-coating procedure for the low-loss sodium film. a, Optical images of the procedure. **b**, Atomic force microscope topology of the smooth quartz surface, which shows that the root mean square (RMS) roughness is about 100 pm.



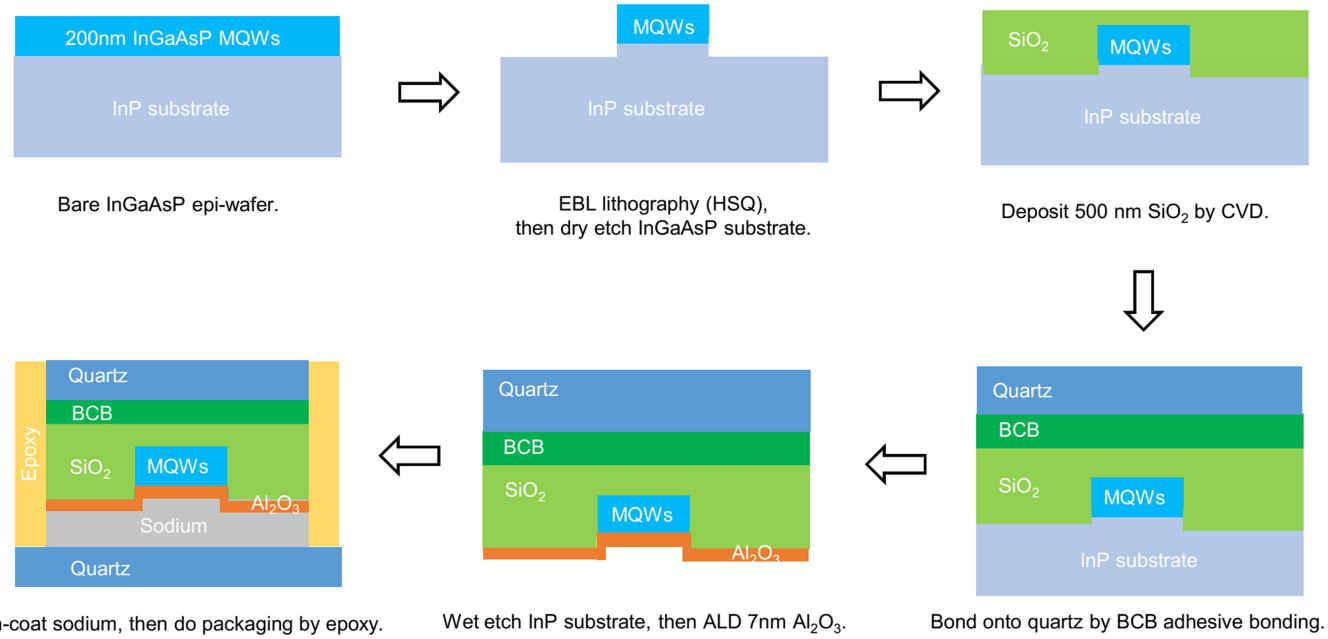
Extended Data Fig. 2 | Dielectric function measurements for sodium and comparison with silver. a, Schematic of ellipsometer measurements. The incident light transmits through the upper surface to the lower surface of the silica layer (contact with the sodium interface) and reflects back. The light reflected from the sodium/glass interface can be detected. **b**, Comparison of real (**a**) and imaginary (**b**) parts as well as the figure of merit ($-\varepsilon_1/\varepsilon_2$) (**c**) of the dielectric functions of silver and sodium (this work). The data for silver are taken from the literature[17,18,23,24,34–37].



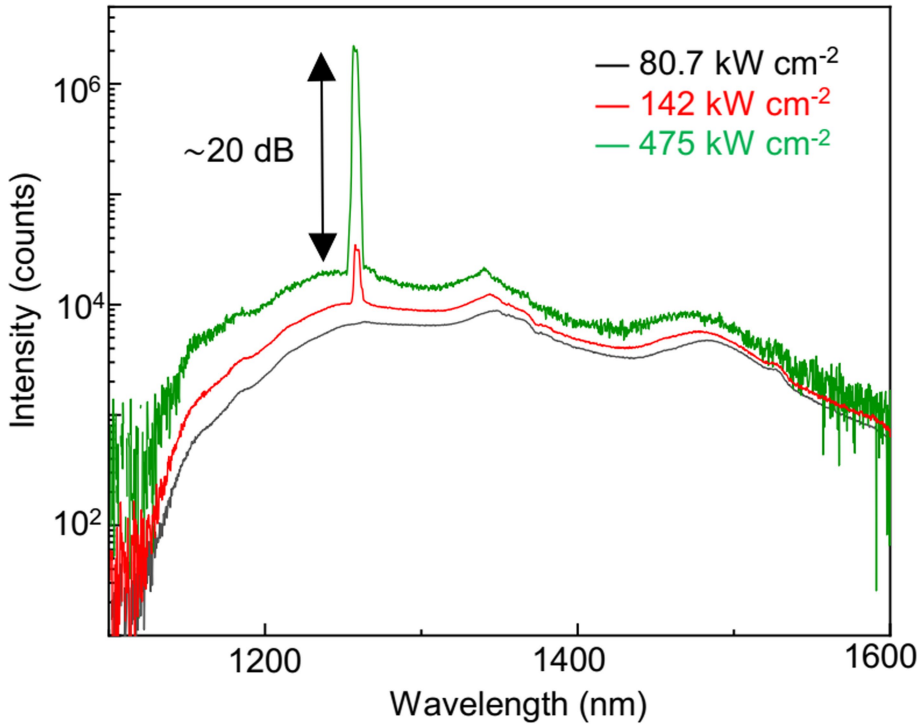
Extended Data Fig. 3 | Detailed fabrication process of the plasmonic waveguides. FIB, focused ion beam.



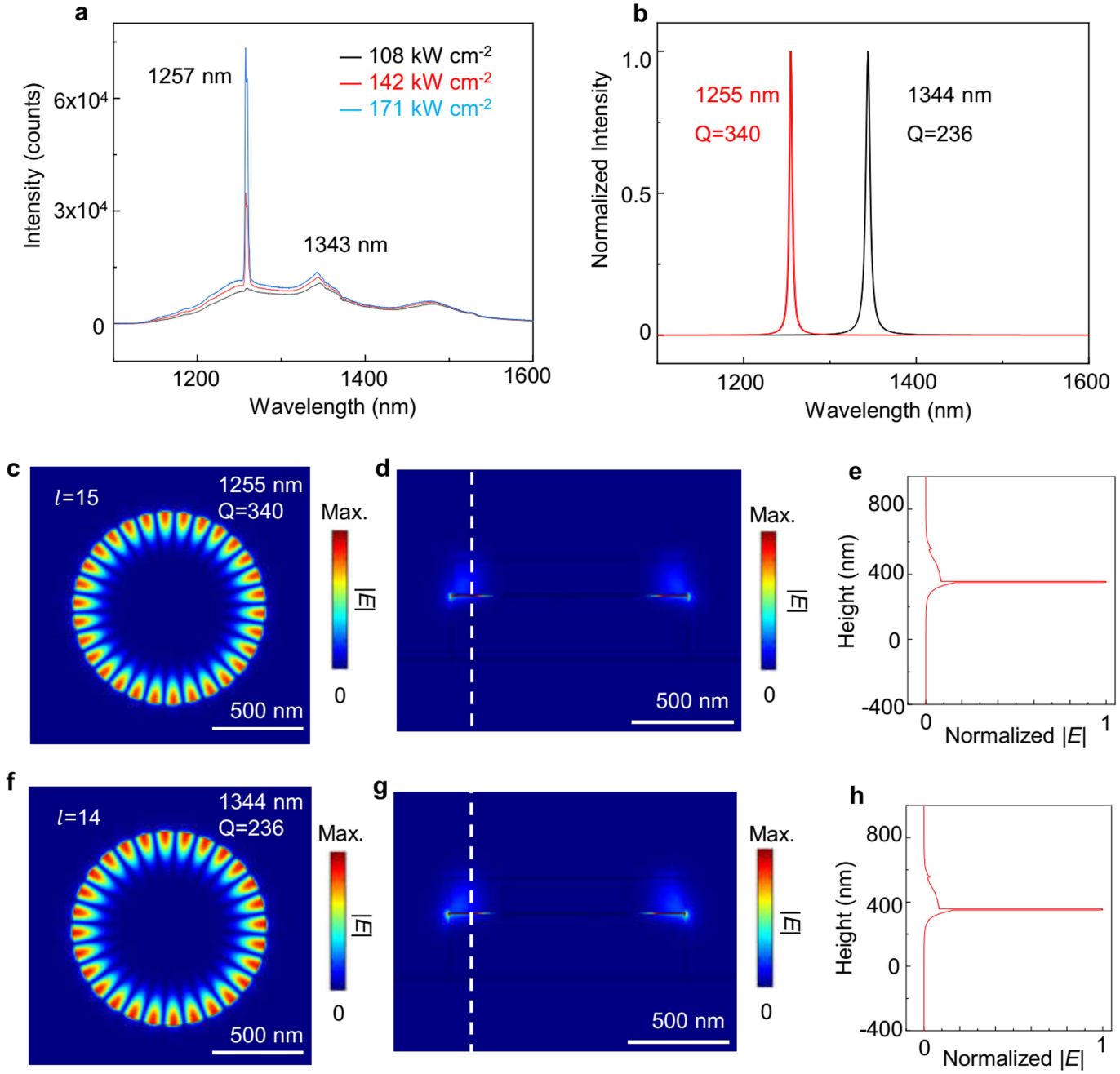
Extended Data Fig. 4 | Comparison of silver and sodium plasmonic waveguides. a, SEM image of the launching and outcoupling structures of plasmonic waveguides. **b**, The fitted δ_{SPP} at different wavelengths on the quartz interface. The dashed line is a fit to the data. **c**, Propagation measurements at wavelengths of 1,180 nm and 1,450 nm for sodium and silver, with the exponential curves fitted to the data. **d**, Figures of merit for sodium to silver plasmonic waveguides. The figure of merit of the plasmonic waveguides is defined as the ratio of propagation length to skin depth $\delta_{SPP}/(\delta_d + \delta_m)$. The sodium data are calculated from the measured dielectric function, and the silver data are calculated using the dielectric function, taken from a number of representative publications[17,18,23,24,34–37].



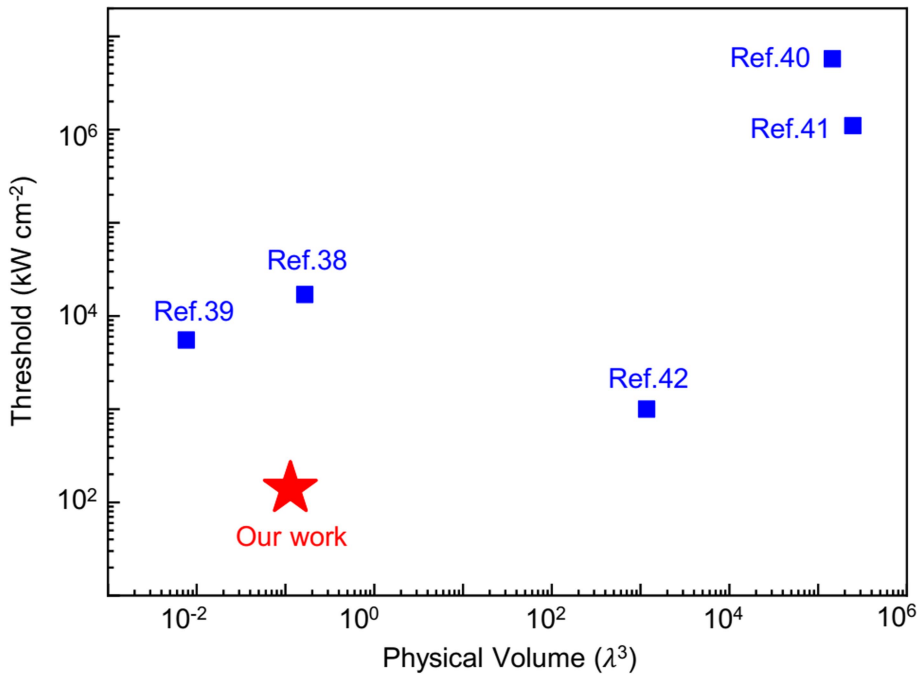
Extended Data Fig. 5 | Detailed fabrication process of the sodium-based plasmonic nanolaser. BCB, benzocyclobutene; HSQ, hydrogen silsesquioxane; EBL, electron-beam lithography; ALD, atomic layer deposition; CVD, chemical vapour deposition.



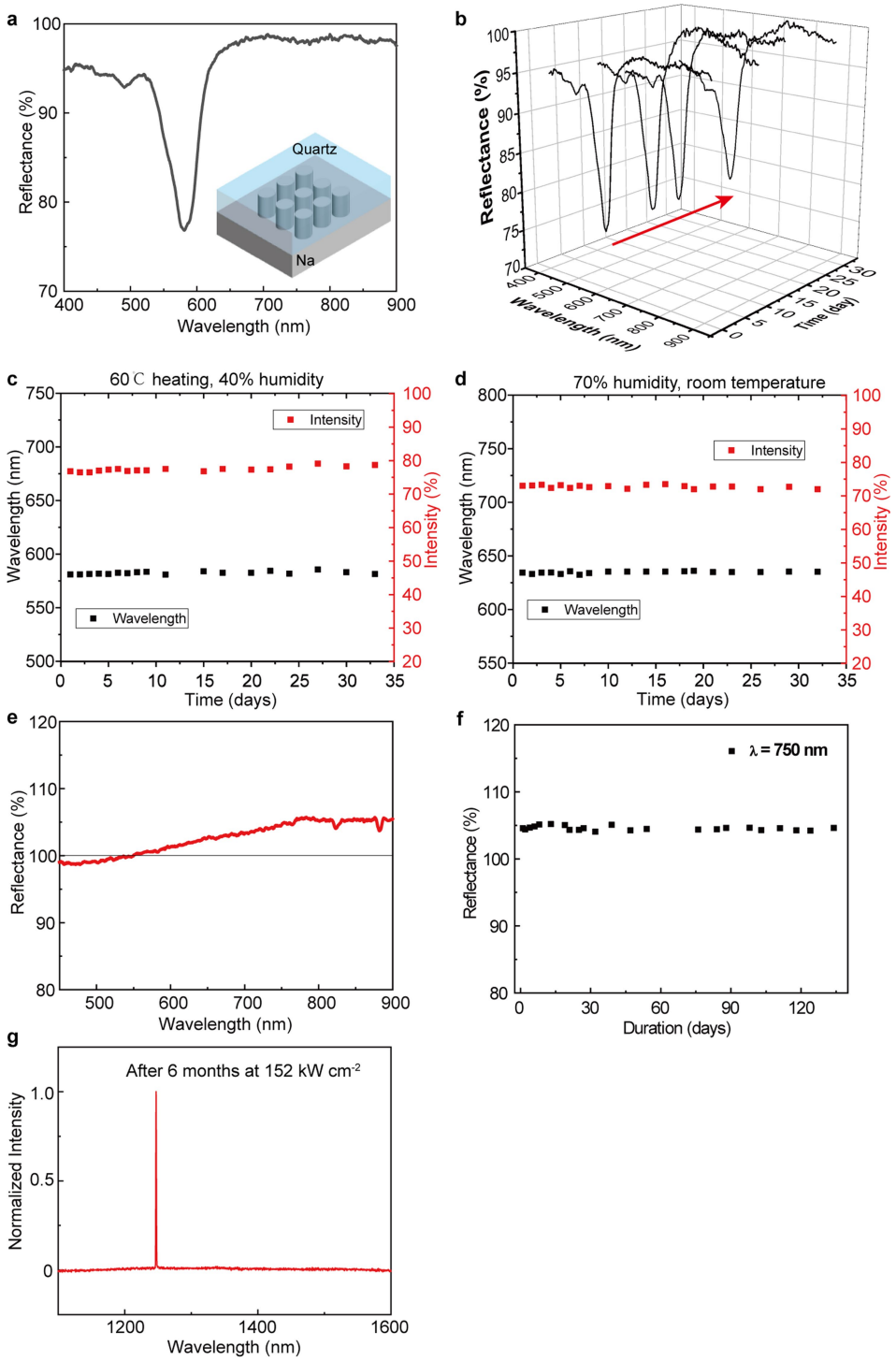
Extended Data Fig. 6 | Emission spectra (semi-logarithmic scale) of the plasmonic nanolaser shown in the main text for different pump powers. When the pump intensity is 475 kW cm^{-2}, the device shows single-mode lasing with a side-mode suppression ratio of 20 dB.



Extended Data Fig. 7 | Numerical simulation of the cavity modes of the sodium-based plasmonic nanolaser. a, Measured emission spectra of the plasmonic nanolaser around the lasing threshold, which exhibit two clear cavity modes at wavelengths of 1,257 nm and 1,343 nm respectively. **b**, Spectrum profile of the two cavity modes, where the resonant wavelength and the linewidth are obtained from the simulation. We assumed Lorentz-shape spectrum profiles here. The simulation shows that the cavity modes at wavelengths of 1,257 nm and 1,343 nm can be identified as plasmonic whispering-galley modes with azimuthal orders of 15 and 14, respectively, and the corresponding quality factors Q of these two modes are 340 and 236, respectively. **c**, **d**, Top (**c**) and side (**d**) views of simulated electric-field distributions of the plasmonic whispering-galley modes with azimuthal order of 15. **e**, Electric-field profile E extracted from the white dashed line in **d**. **f**, **g**, Top (**f**) and side (**g**) views of simulated electric-field distributions of the plasmonic whispering-galley modes with azimuthal order of 14. **h**, Electric-field profile extracted from the white dashed line in **g**.



Extended Data Fig. 8 | Comparison of our work to reported plasmonic lasers in terms of the threshold and cavity size. Blue square symbols represent reported plasmonic nanolasers in the literature. The red star represents our sodium-based plasmonic nanolaser. λ is the lasing emission wavelength in free space.



Extended Data Fig. 9 | Stability test for sodium-based reflective plasmonic filters, sodium mirror and plasmonic nanolasers. a–**d**, Accelerated ageing tests for sodium-based reflective plasmonic filters. **a**, Reflectance spectrum of a representative filter device. Inset, schematic of the device. **b**, Reflectance spectra over time. **c**, Stability of a fabricated device in terms of the resonance wavelength and intensity at 60 °C, with air humidity 40%. **d**, Stability of a fabricated device in terms of the resonance wavelength and intensity at room temperature and air humidity 70%. **e**, **f**, Stability of the sodium mirror. **e**, Reflectance spectrum of the sodium mirror against a standard silver mirror (Thorlabs, PF10-03-P01). **f**, Reflectance of the packaged sodium mirror at wavelength $\lambda = 750$ nm over 120 days in air. **g**, Lasing spectrum of a sodium-based plasmonic nanolaser after six months.

Extended Data Table 1 | Performance comparison of near-infrared plasmonic nanolasers

Table 1|Performance comparison of near-infrared plasmonic nanolasers

Year	Type	Feedback	Size	Gain medium	Emission Wavelength	Pump	Temperature	Threshold	Refs
2020	Plasmonic nanodisk	WGM	Diameter: 1.2 μm Height: 200 nm	InGaAsP	1257 nm	Optical, pulsed	Room Temperature	140 kW cm^{-2}	Our work
2009	Gap plasmon	FP	Thickness: 310 nm Length: 6 μm	InGaAs	1500 nm	Electrical, continuous wave	Room Temperature	17 MW cm^{-2}	38
2012	Plasmonic coaxial	FP	Diameter: 500 nm Height: 260 nm	InGaAsP	1400 nm	Optical, continuous wave /pulsed	Room Temperature	5.5 MW cm^{-2}	39
2013	Plasmonic nanocavity array	Bandedge lattice plasmons	Diameter: 130 nm Period: 600 nm Height: 60 nm	Dye	913 nm	Optical, pulsed	Room temperature	5.75 GW cm^{-2}	40
2015	Plasmonic nanocavity array	Bandedge lattice plasmons	Diameter: 120 nm Period: 600 nm Depth: 150 nm	Dye	855 nm	Optical, pulsed	Room temperature	1.1 GW cm^{-2}	41
2018	Plasmonic DFB	FP	Length: 2 mm Width: 1 μm Height: 450 nm	Dye	800 nm	Optical, pulsed	Room Temperature	1 MW cm^{-2}	42
2010	Plasmonic nanodisk	WGM	Thickness: 235 nm Diameter: 1 μm	InAsP	1300 nm	Optical, pulsed	8 K	120 kW cm^{-2}	43
2011	Plasmonic DFB	FP	Width: 140 nm Length: 4 μm	InGaAs	1550 nm	Electrical, continuous wave	80 K	134 kW cm^{-2}	44
2011	Plasmonic nanodisk	WGM	Diameter: 1.25 μm Height: 235 nm	InP	1300 nm	Optical, pulsed	8 K	26 kW cm^{-2}	45
2011	Plasmonic crystal defect	FP	Width: 380 nm Length: 14 μm	InGaAsP	1350 nm	Optical, pulsed	77 K	100 kW cm^{-2}	46
2015	Plasmonic nanowire	FP	Diameter: 150 nm Length: 4.77 μm	AlGaAs/ GaAs	800 nm	Optical, pulsed	8 K	4 MW cm^{-2}	47
2016	Plasmonic nanowire	FP	Diameter: 180 nm Length: 2.3 μm	GaAs	850 nm	Optical, pulsed	7 K	2.4 MW cm^{-2}	48
2017	Plasmonic nanowire	FP	Length: 4.8 μm Width: 380 nm Height: 140 nm	InGaAsP	1300 nm	Optical, continuous wave	125 K	1.4 MW cm^{-2}	49

DFB, distributed feedback; WGM, whispering gallery mode; FP, Fabry-Perot.

Seeded growth of large single-crystal copper foils with high-index facets

https://doi.org/10.1038/s41586-020-2298-5

Received: 29 March 2019

Accepted: 11 March 2020

Published online: 27 May 2020

Check for updates

Muhong Wu[1,2,3,16], Zhibin Zhang[1,16], Xiaozhi Xu[4,16], Zhihong Zhang[2,16], Yunrui Duan[5,6], Jichen Dong[5], Ruixi Qiao[2], Sifan You[2], Li Wang[1,7], Jiajie Qi[1], Dingxin Zou[8], Nianze Shang[1], Yubo Yang[9], Hui Li[6], Lan Zhu[10], Junliang Sun[11], Haijun Yu[9], Peng Gao[2], Xuedong Bai[7], Ying Jiang[2,12], Zhu-Jun Wang[13], Feng Ding[5,14]✉, Dapeng Yu[8]✉, Enge Wang[2,3,15]✉ & Kaihui Liu[1,2]✉

The production of large single-crystal metal foils with various facet indices has long been a pursuit in materials science owing to their potential applications in crystal epitaxy, catalysis, electronics and thermal engineering[1–5]. For a given metal, there are only three sets of low-index facets ({100}, {110} and {111}). In comparison, high-index facets are in principle infinite and could afford richer surface structures and properties. However, the controlled preparation of single-crystal foils with high-index facets is challenging, because they are neither thermodynamically[6,7] nor kinetically[3] favourable compared to low-index facets[6–18]. Here we report a seeded growth technique for building a library of single-crystal copper foils with sizes of about 30 × 20 square centimetres and more than 30 kinds of facet. A mild pre-oxidation of polycrystalline copper foils, followed by annealing in a reducing atmosphere, leads to the growth of high-index copper facets that cover almost the entire foil and have the potential of growing to lengths of several metres. The creation of oxide surface layers on our foils means that surface energy minimization is not a key determinant of facet selection for growth, as is usually the case. Instead, facet selection is dictated randomly by the facet of the largest grain (irrespective of its surface energy), which consumes smaller grains and eliminates grain boundaries. Our high-index foils can be used as seeds for the growth of other Cu foils along either the in-plane or the out-of-plane direction. We show that this technique is also applicable to the growth of high-index single-crystal nickel foils, and we explore the possibility of using our high-index copper foils as substrates for the epitaxial growth of two-dimensional materials. Other applications are expected in selective catalysis, low-impedance electrical conduction and heat dissipation.

At present, high-index metal foils are mainly obtained by cutting bulk single-crystal metal ingots or by epitaxial deposition on other non-metal single crystals with a high-index facet[19,20]. Thus, only a few index choices are available, and the cost is very high. Additionally, in the typical cutting and polishing method, accurate index control and flat surfaces with uniformly parallel step edges are very challenging to realize (with typically a 1°–3° tolerance in the cutting angle of single crystals from commercial products). Hence, new methods that enable efficient production of large single-crystal metal foils with various high-index facets are in great demand.

Recently, the synthesis of noble metal nanocrystals with high-index facets by electrochemical reaction, electrodeposition or solution-phase synthesis has been reported[3,4]. The basic idea in high-index noble metal nanocrystal growth is to perturb the thermodynamic equilibrium state of a low-index facet or to modify the kinetic barrier to the growth of a high-index seed. In our experiment, we first developed a pre-oxidation treatment technique to generate a large grain seed with a high-index facet in a commercial polycrystalline Cu foil. The Cu foils were oxidized at 150–650 °C in air, typically for a few hours, and then annealed in a reducing atmosphere at 1,020 °C for

[1]State Key Laboratory for Mesoscopic Physics, Frontiers Science Center for Nano-optoelectronics, School of Physics, Peking University, Beijing, China. [2]International Center for Quantum Materials, Collaborative Innovation Center of Quantum Matter, Peking University, Beijing, China. [3]Songshan Lake Materials Laboratory, Institute of Physics, Chinese Academy of Sciences, Guangdong, China. [4]Guangdong Provincial Key Laboratory of Quantum Engineering and Quantum Materials, School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou, China. [5]Centre for Multidimensional Carbon Materials, Institute for Basic Science, Ulsan, South Korea. [6]Key Laboratory for Liquid–Solid Structural Evolution and Processing of Materials, Ministry of Education, Shandong University, Jinan, China. [7]Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing, China. [8]Shenzhen Institute for Quantum Science and Engineering, and Department of Physics, Southern University of Science and Technology, Shenzhen, China. [9]College of Materials Science and Engineering, Beijing University of Technology, Beijing, China. [10]Peking Union Medical College Hospital, Beijing, China. [11]College of Chemistry and Molecular Engineering, Peking University, Beijing, China. [12]Physical Science Laboratory, Huairou National Comprehensive Science Centre, Beijing, China. [13]Scientific Centre for Optical and Electron Microscopy, Eidgenössische Technische Hochschule Zürich, Zurich, Switzerland. [14]School of Materials Science and Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea. [15]School of Physics, Liaoning University, Shenyang, China. [16]These authors contributed equally: Muhong Wu, Zhibin Zhang, Xiaozhi Xu, Zhihong Zhang. ✉e-mail: f.ding@unist.ac.kr; yudp@sustc.edu.cn; egwang@pku.edu.cn; khliu@pku.edu.cn

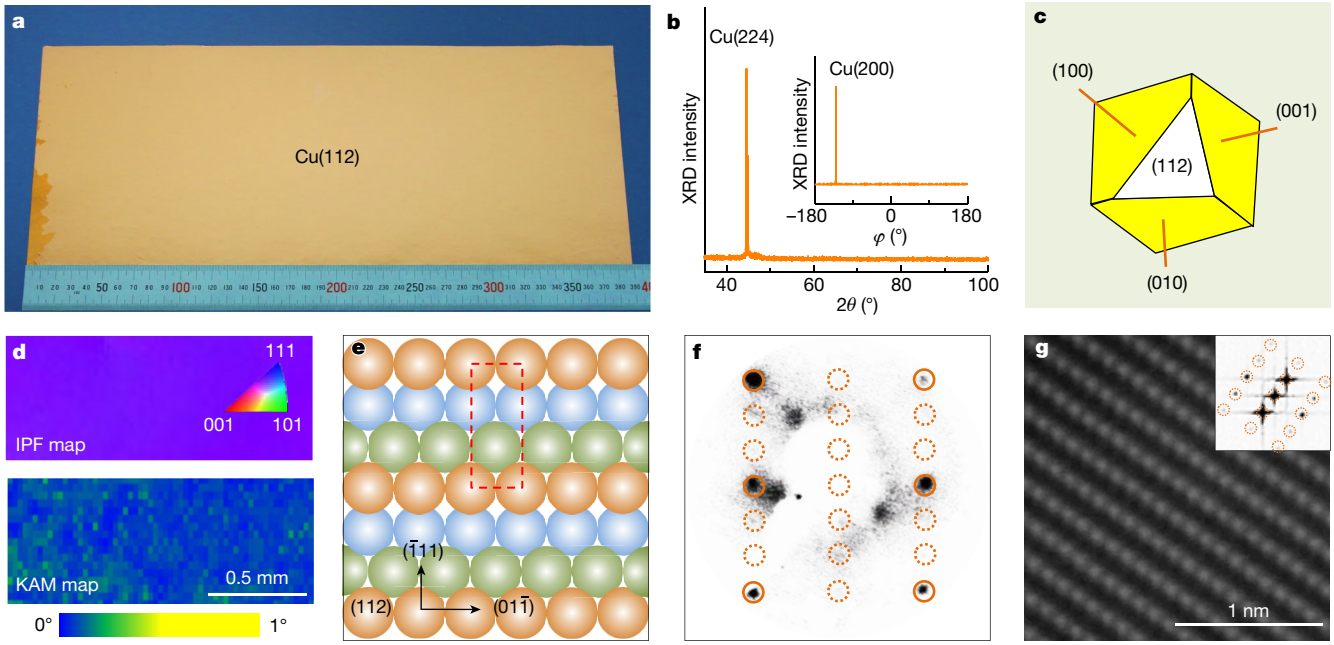


Fig. 1 | Characterization of a Cu(112) single crystal annealed from a commercial polycrystalline foil. a, Optical image of a single-crystal Cu(112) after mild oxidation in air at 200 °C. The single-crystal Cu(112) shows a homogeneous colour (the small darker pattern at the bottom-left corner is the polycrystalline part). The foil size is 39 × 21 cm^2. The rulers shown in all figures indicate millimetres unless stated otherwise. **b**, XRD 2θ scan data for the annealed Cu foil with (224) or (112) orientation. Inset, azimuthal off-axis φ scan spectrum with only one peak corresponding to Cu(200), confirming single-crystal features without in-plane rotation. **c**, Reconstructed single-crystal XRD pattern of the Cu(112) foil. **d**, Representative IPF (top) and KAM (bottom) maps of a Cu(112) foil. **e**, Schematic diagram of the Cu(112) surface. The orange, green and blue balls correspond to the first, second and third layers of Cu atoms from the surface, respectively. The red dashed box indicates the primitive cell for the first layer of Cu(112). **f**, Representative LEED pattern of the as-prepared Cu(112) foil. **g**, Atomically resolved STEM image of the Cu(112) foil. Inset, fast Fourier transform pattern of the STEM image.

several hours (see Methods for more details). The as-annealed Cu foil was heated in air at about 200 °C for 2–30 min before optical imaging. Owing to the difference in oxidation barriers, different types of facets show distinct colours, and a single-crystal Cu domain can be clearly observed as a homogeneous region in an optical image[21]. After thermal annealing, the whole ~39 × 21 cm^2 commercial polycrystalline Cu foil was transformed into a large single crystal (see Fig. 1a, where only the small dark orange zone at the bottom-left corner is polycrystalline), which is the largest single-crystal Cu foil reported so far, to our knowledge. This appealing property is enabled by our large annealing furnace, with a heating area of length 50 cm and diameter 23 cm, which provides a highly stable temperature distribution. The X-ray diffraction (XRD) 2θ scan (Fig. 1b) reveals that after pre-oxidation and subsequent annealing, the facet of the single crystal is not Cu(111), as reported before[12–14], but Cu(224) or Cu(112) (only facets with all Miller indices h, k and l even or odd give a non-zero 2θ scan signal in XRD owing to the extinction rule). A further azimuthal off-axis φ scan shows only one peak corresponding to Cu(200) (inset in Fig. 1b), demonstrating that the as-annealed Cu foil is a single crystal without in-plane rotation. To further confirm its crystal structure, single-crystal XRD was performed, and lattice reconstruction using a standard procedure confirms the (112) facet (Fig. 1c). Electron back-scattered diffraction (EBSD) was also applied to analyse the surface crystallography. The uniform violet colour in the inverse pole figure (IPF) map verifies the (112) facet index (top panel in Fig. 1d); the kernel average misorientation (KAM) map (bottom panel in Fig. 1d) shows that the local average misorientation between each measured point and its nearest neighbours is extremely small (typically less than 0.3°), indicating that the as-annealed Cu foil is a homogeneous single crystal. The (112) lattice structure of the as-annealed Cu (schematic in Fig. 1e) was also demonstrated in both reciprocal space and real space by low-energy electron diffraction (LEED; Fig. 1f) and scanning transmission electron microscopy (STEM; Fig. 1g), respectively.

With our design, the Cu(111) facet is no longer the only thermodynamically favourable facet, and a high-index facet, such as Cu(112), can be successfully synthesized. We repeated the procedure on many Cu foils, and more than 30 kinds of high-index single crystal with large domain sizes were prepared (length 25–39 cm, width 21 cm). Eight representative kinds of Cu foil with different facet indices and a typical size of 35 × 21 cm^2 are shown in Fig. 2a (the large uniform regions are single crystals). The single-crystal nature of the annealed Cu foils was further confirmed by the distinct colours in the corresponding EBSD IPF maps (Fig. 2b). The facet indices were identified by the characteristic peaks in the XRD 2θ scan spectra (Fig. 2c) and reconstructed single-crystal XRD data (Extended Data Fig. 1). Some high-index surface structures were further revealed by atom-resolved scanning tunnelling microscopy (STM) images (Extended Data Fig. 2), in which the spacing between atomic stripes matches the theoretical values from the atomic model.

To verify the essential role of the pre-oxidation treatment in producing high-index facets (Fig. 3a, stage 1), we carried out a control experiment in which we locally oxidized one end of a piece of polycrystalline Cu foil while the other end remained intact (Fig. 3b). After the annealing procedure, the pre-oxidized part was transformed into Cu(235) and the untreated part was transformed into Cu(111) (Fig. 3c, Extended Data Fig. 3). Statistical data further revealed that only Cu(111) foils were obtained without pre-oxidation; when the Cu foils were pre-oxidized, more than 30 types of facet were observed (Fig. 3d, e, Extended Data Table 1).

The above observations clearly reveal that pre-oxidation must induce some critical factor that controls the formation of high-index seeds. It is well established that the stored strain energy and the surface energy in a thin metal film are competitive driving forces for abnormal grain growth during annealing, with the former leading to abnormal Cu(001) grains because the smallest biaxial modulus of Cu is along the ⟨001⟩ direction[9], and the latter giving the facet with the minimum surface energy. Before the start of the abnormal grain growth, our thin (only

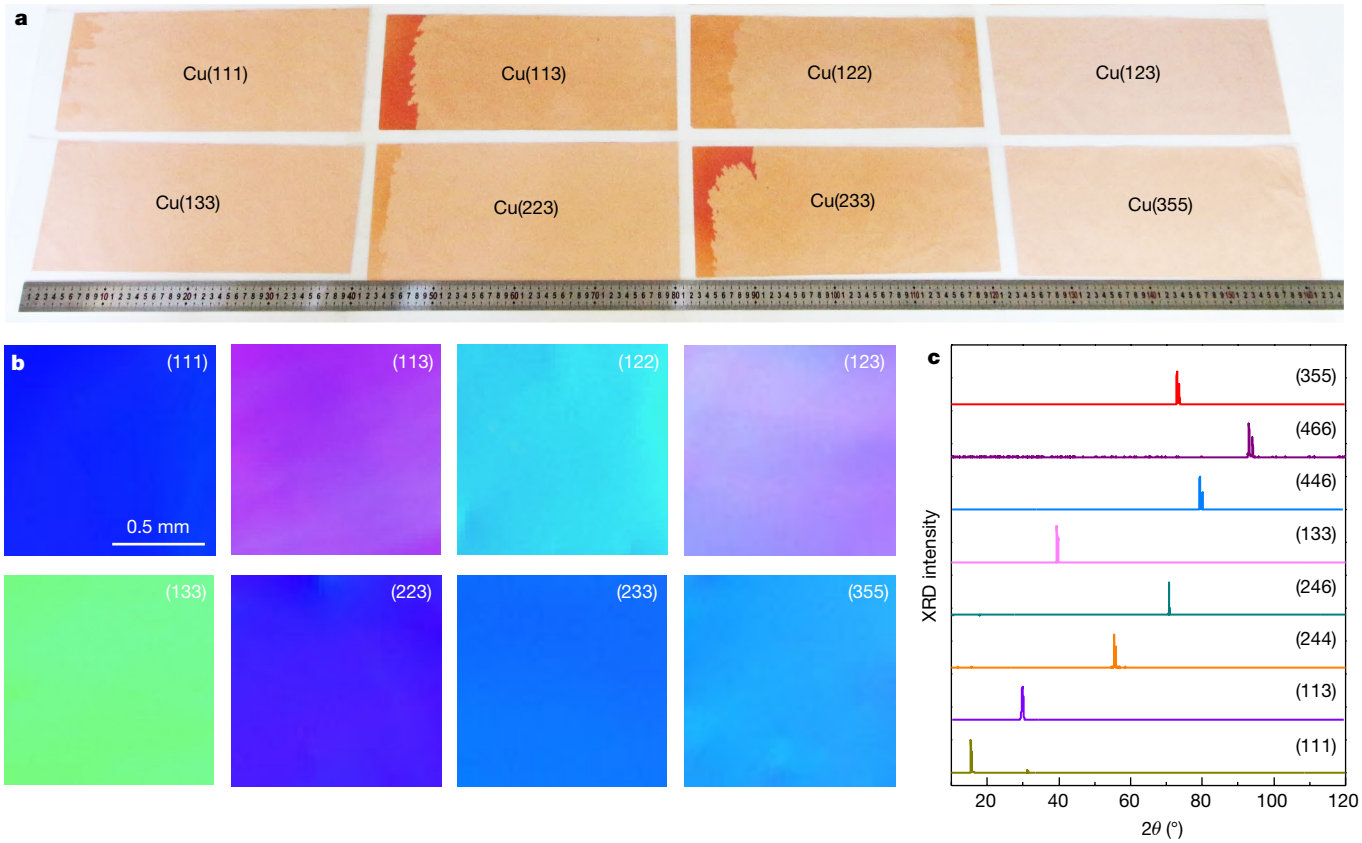


Fig. 2 | Single-crystal Cu foils of A4 paper size with various facet indices. **a**, Optical image of the eight representative types of single-crystal Cu foil with typical size of 35 × 21 cm^2. A mild oxidation treatment was performed to quickly identify the single-crystal region in the foils (the uniform areas are single crystals; the coloured corners are polycrystalline). The ruler indicates centimetres. **b**, EBSD IPF maps of the eight kinds of single-crystal Cu foil. The distinct colours reveal their different facet indices. These maps have the same size. **c**, XRD 2θ scan spectra of the eight kinds of single-crystal Cu foil. The characteristic peaks are used to identify the facet indices (owing to the extinction rule in XRD, only peaks corresponding to (244), (246), (446) and (466) appear, respectively, for the (122), (123), (223) and (233) facets). The X-ray source is a silver-based target with wavelength of about 0.56 Å and all the peaks are normalized with respect to the intensity maximum.

25 μm thick) Cu foil was annealed at 1,020 °C for a long time; thus, most of the stored strain energy was released, and the abnormal grain growth was mainly driven by the surface energy. As a consequence, annealing of an original Cu foil results in a Cu(111) single crystal with the lowest surface energy among all possible facets[12–14].

On the other hand, when Cu is mildly oxidized, both of its upper and lower surfaces are covered by a layer of Cu_xO grains with different orientations (Extended Data Fig. 4). Thus, the two free surfaces of a Cu grain are transformed into two Cu–Cu_xO interfaces after pre-oxidation, and then the interface energy, which is the summation of the energies of all types of interfaces between many small polycrystalline Cu_xO grains and the Cu foil, depends weakly on the orientation of the Cu grain itself. Therefore, grain seeds with various surface indices (Cu(*hkl*)) have certain probabilities of growing abnormally until the oxide layer is completely reduced. With further annealing, the Cu(*hkl*) grain seed with an advantage in size spreads to the whole Cu foil by abnormal grain growth (Fig. 3b, stage 2).

Unlike abnormal grain seeding, which is mainly driven by thermodynamics, abnormal grain growth is mainly controlled by kinetics. The large abnormal grain consumes the surrounding small normal grains and eliminates the grain boundaries in the Cu foil. In such a process, excessive oxygen at the grain boundaries acts as pinning centres and lowers the mobility of the grain boundaries[22]. Introducing hydrogen during annealing removes excessive oxygen from the Cu foil and accelerates the abnormal grain growth. In addition, the designed in-plane temperature difference throughout the Cu foil facilitates only one abnormal grain acting as a seed in the high-temperature region in most annealing processes (Extended Data Fig. 5). We also notice that the surface energy of pure Cu may play some role during the abnormal grain growth, favouring the formation of low-energy facets, because some parts of the oxide layer could be reduced to Cu during the annealing in the abnormal grain seeding stage[23–30] (Fig. 3f, Extended Data Table 1).

Using this technique, we can readily obtain large single-crystal Cu foils with various high-index facets. However, the appearance of a certain high-index seed is a random event and cannot be experimentally designed. To controllably produce the desired high-index single-crystal Cu foils, we further propose a facet 'transfer' method to copy the facet of the obtained high-index single crystals. First, a small piece of high-index single-crystal Cu, cut from a large single-crystal Cu foil obtained by pre-oxidation annealing, was placed on a large polycrystalline Cu foil to serve as a new seed. When annealed at a high temperature (but below the melting point of the bulk Cu), the surface of the seed started to assimilate into the polycrystalline Cu, and the single-crystal lattice arrangement had a very high chance (>98% under our current experimental conditions) of transferring to the polycrystalline Cu, resulting in a 'nucleus' with the same facet as the seed (Extended Data Fig. 6a). The time evolution of the thermal annealing process verifies that the facet 'transfer' indeed works, as shown by the example Cu(245) single crystal. The triangular single-crystal Cu(245) piece first assimilated from the top surface, on which the seed was placed, to the bottom surface of the polycrystalline Cu foil (Fig. 4a, b), and a new Cu(245) seed was formed in the polycrystalline foil that then spread into a large single crystal (Fig. 4c, d). The identical colour of the EBSD IPF maps (Fig. 4e), together with the XRD 2θ and φ scan data for the initial seed

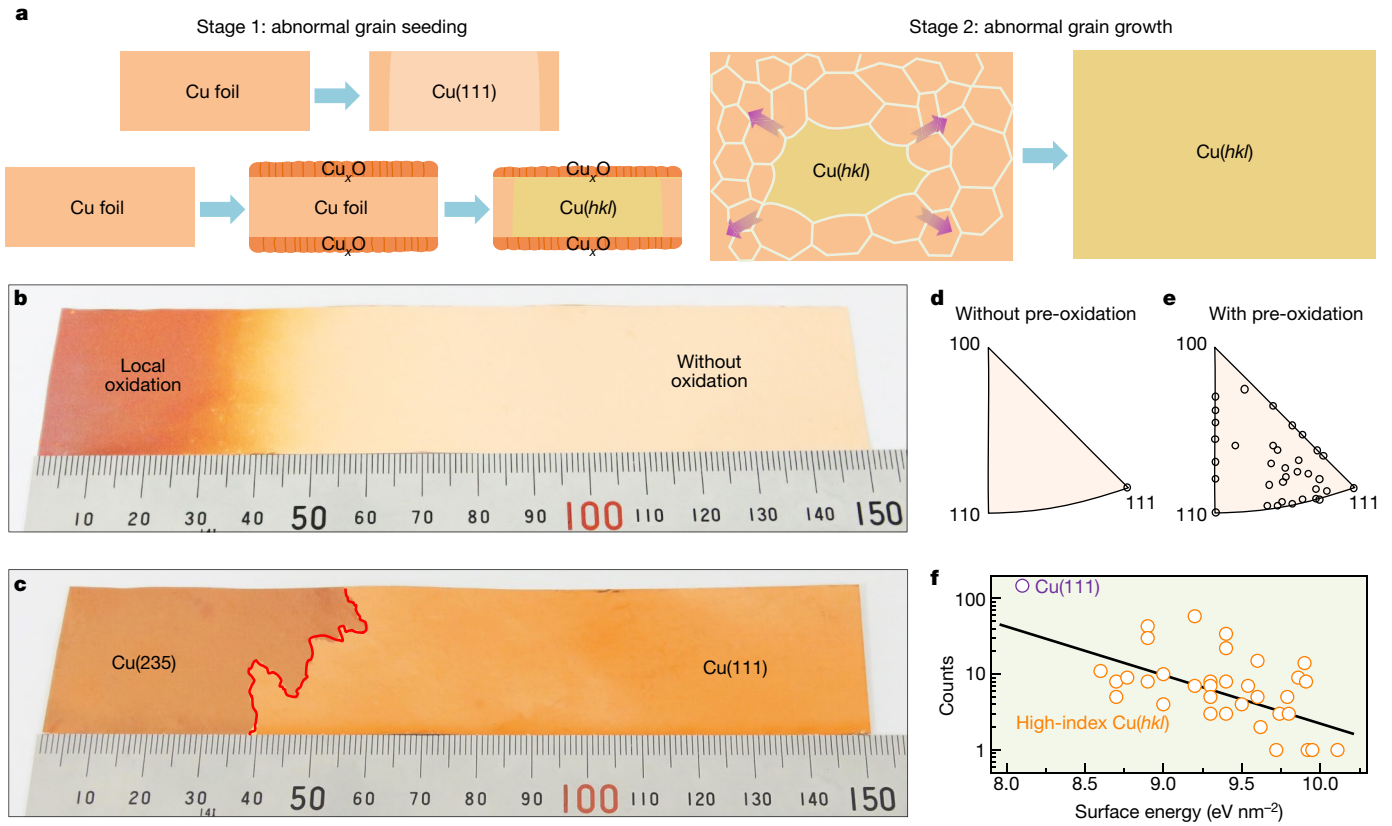


Fig. 3 | Pre-oxidation-guided seeded growth of single-crystal Cu foils with high facet indices. a, Schematic diagrams of the two stages of the fabrication of high-index single crystals. Stage 1: without an oxide layer, Cu(111) has the lowest surface energy and is the prevailing growing facet. Once the oxide layer is formed, high-index Cu(*hkl*) can grow, and form the abnormal grain seed. Stage 2: the large abnormal grain seed consumes the surrounding small normal grains and eliminates the grain boundaries in the Cu foil. **b**, **c**, Optical images of a Cu foil before (**b**) and after (**c**) annealing. One terminal of the Cu foil was pre-oxidized locally and the other one was kept intact before annealing. After annealing, the pre-oxidized terminal transformed into Cu(235) and the intact one changed into Cu(111). **d**, **e**, IPFs in the normal direction for the distribution of facet indices of Cu single crystals annealed without (**d**) and with (**e**) pre-oxidation. 35 kinds of Cu facet appeared with the pre-oxidation treatment, in striking contrast to the occurrence of only the Cu(111) facet without pre-oxidation. **f**, Appearance frequency of facets as a function of their surface energy for the single-crystal Cu foils obtained. The general trend shows that facets with lower surface energies can be achieved more easily. All the samples were pre-oxidized under the same conditions (150 °C for 2 h) and had almost the same oxide thickness (~50 nm).

and the final obtained Cu (Extended Data Fig. 6b–f), showed that the facet of the seed was successfully copied and a new large-size Cu(245) single crystal was produced. In addition to this in-plane facet transfer, we also found that a Cu foil can be transformed into 'bulk' single-crystal Cu by an out-of-plane facet transfer process, as shown in Fig. 4f. Here, the length of each foil was deliberately increased by about 2 mm with respect to the previous one to facilitate characterization of the obtained single crystal. The identical colour of the EBSD IPF maps for the different layers of Cu foils after cyclic facet 'transfer', together with the XRD characterization for the whole thick Cu foil, demonstrated that the single-crystal structure of the first Cu(256) layer was successfully transferred throughout the foil (Fig. 4g–i, Extended Data Fig. 6g), revealing a new facile way to prepare a bulk single-crystal Cu slab with controlled crystal orientation. The above processes of vertical assimilation and in-plane growth were explained and reproduced by our molecular dynamics simulations (Extended Data Fig. 7).

To demonstrate the universality of this seeded abnormal grain growth technique, we attempted to anneal nickel (Ni) using a similar approach at about 1,200 °C. Three types of high-index single-crystal Ni foil with (012), (013) and (355) facets, typically with a size of about 5 × 5 cm^2, were obtained (Fig. 4j–n, Extended Data Fig. 8). The Ni single-crystal size was limited by the diameter of the alundum tubes in our custom-designed high-temperature chemical vapour deposition system. In addition to Cu and Ni, we expect that similar approaches could be applied to grow single-crystal foils of many other metals (such as Fe, Al, Ag and Pt) and their alloys. Moreover, we found that the epitaxial growth of graphene and hexagonal boron nitride (hBN) is also achievable on high-index facets (Extended Data Fig. 9), not only on Cu(111) (refs. [12,13]) or vicinal Cu(110) (ref. [18]), as previously reported. We therefore expect further advancement in the fundamental exploration and technical applications of these large high-index single-crystal metal foils.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2298-5.

1. Lin, C. et al. Two-dimensional wetting of a stepped copper surface. *Phys. Rev. Lett.* **120**, 076101 (2018).
2. Park, C. H. et al. Anisotropic behaviours of massless Dirac fermions in graphene under periodic potentials. *Nat. Phys.* **4**, 213–217 (2008).
3. Tian, N. et al. Synthesis of tetrahexahedral platinum nanocrystals with high-index facets and high electro-oxidation activity. *Science* **316**, 732–735 (2007).
4. Quan, Z. W., Wang, Y. X. & Fang, J. Y. High-index faceted noble metal nanocrystals. *Acc. Chem. Res.* **46**, 191–202 (2013).
5. Huang, X. et al. Solution-phase epitaxial growth of noble metal nanostructures on dispersible single-layer molybdenum disulfide nanosheets. *Nat. Commun.* **4**, 1444 (2013).

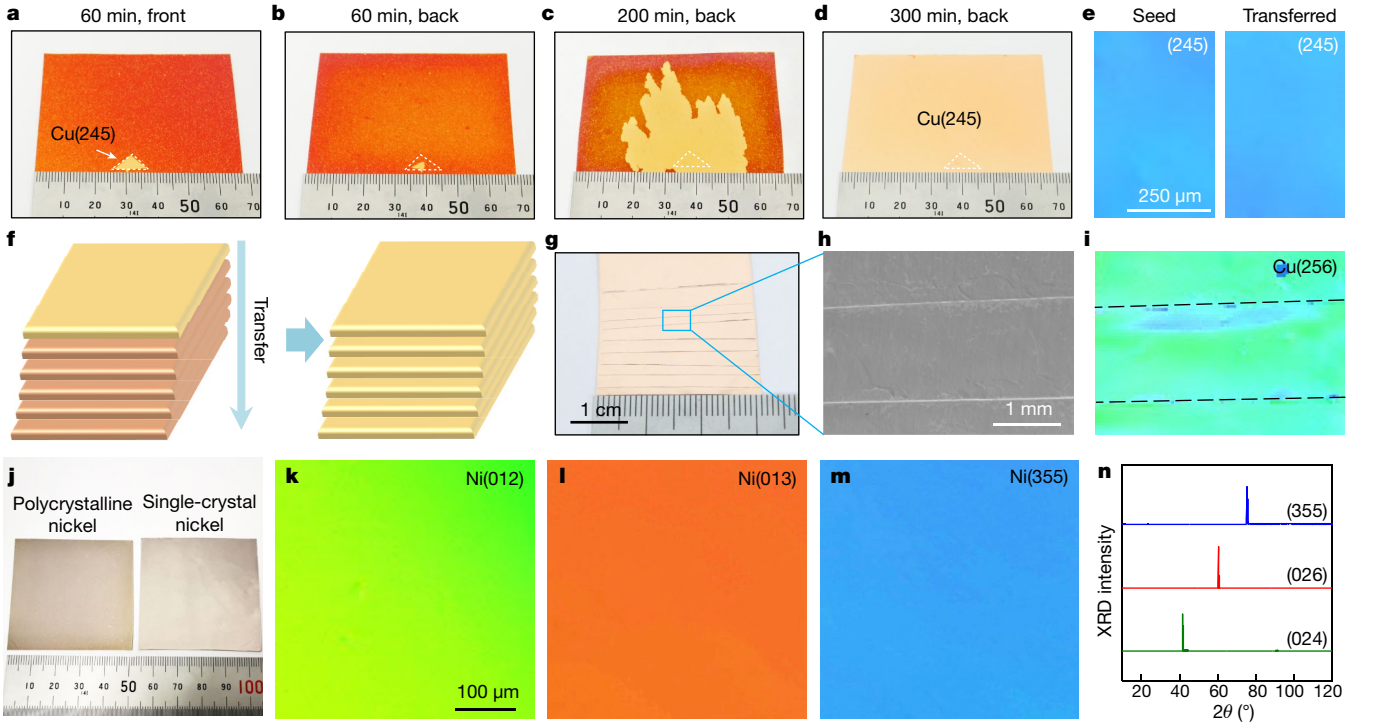


Fig. 4 | Facet transfer of single-crystal Cu and characterization of high-index single-crystal Ni foils. a, Photograph of the front surface of a Cu foil after seed-guided annealing for 60 min. **b**–**d**, Photographs of the back surfaces of Cu foils after seed-guided annealing for 60 min (**b**), 200 min (**c**) and 300 min (**d**). **e**, EBSD IPF maps of the seed (left) and the Cu single crystals with the transferred facet (right). The identical colour shows that the (245) single-crystal structure was copied successfully. The two EBSD maps have the same size. The facet index is determined from the XRD characterization in Extended Data Fig. 6b–f. **f**, Schematic diagrams illustrating out-of-plane facet transfer layer by layer. **g**, Optical image of the ten layers of Cu foils with transferred facets. The length of the foils was deliberately increased by ~2 mm each time to facilitate the sample characterization. **h**, **i**, SEM image (**h**) and EBSD IPF map (**i**) of the Cu foils marked in **g**. The uniform colour in the EBSD IPF map, together with the XRD data in Extended Data Fig. 6g, demonstrates that the single-crystal structure of the first Cu layer has been successfully copied by all the foils in the slab. **j**, Photograph of polycrystalline (left) and single-crystal (right) Ni foils with size of about 5 × 5 cm^2. **k**–**m**, EBSD IPF maps of three kinds of single-crystal Ni foil. The distinct colours reveal their different facet indices. These maps are of the same size. **n**, Powder XRD (silver-based target) 2θ scan spectra of the single-crystal Ni foils shown in **k**–**m**.

6. Vitos, L., Ruban, A. V., Skriver, H. L. & Kollar, J. The surface energy of metals. *Surf. Sci.* **411**, 186–202 (1998).
7. Wen, Y. N. & Zhang, H. M. Surface energy calculation of the fcc metals by using the MAEAM. *Solid State Commun.* **144**, 163–167 (2007).
8. Carpenter, H. C. H. & Elam, C. F. The production of single crystals of aluminium and their tensile properties. *Proc. R. Soc. Lond. A* **100**, 329–353 (1921).
9. Zielinski, E. M., Vinci, R. P. & Bravman, J. C. The influence of strain-energy on abnormal grain-growth in copper thin-films. *Appl. Phys. Lett.* **67**, 1078–1080 (1995).
10. Zhang, J. M., Xu, K. W. & Ji, V. Strain-energy-driven abnormal grain growth in copper films on silicon substrates. *J. Cryst. Growth* **226**, 168–174 (2001).
11. Omori, T. et al. Abnormal grain growth induced by cyclic heat treatment. *Science* **341**, 1500–1502 (2013).
12. Brown, L. et al. Polycrystalline graphene with single crystalline electronic structure. *Nano Lett.* **14**, 5706–5711 (2014).
13. Xu, X. Z. et al. Ultrafast epitaxial growth of metre-sized single-crystal graphene on industrial Cu foil. *Sci. Bull.* **62**, 1074–1080 (2017).
14. Jin, S. et al. Colossal grain growth yields single-crystal metal foils by contact-free annealing. *Science* **362**, 1021–1025 (2018).
15. Kusama, T. et al. Ultra-large single crystals by abnormal grain growth. *Nat. Commun.* **8**, 354 (2017).
16. Shan, Z. W. et al. Mechanical annealing and source-limited deformation in submicrometre-diameter Ni crystals. *Nat. Mater.* **7**, 115–119 (2008).
17. Wang, H. et al. Surface monocrystallization of copper foil for fast growth of large single-crystal graphene under free molecular flow. *Adv. Mater.* **28**, 8968–8974 (2016).
18. Wang, L. et al. Epitaxial growth of a 100-square-centimetre single-crystal hexagonal boron nitride monolayer on copper. *Nature* **570**, 91–95 (2019).
19. Uecker, R. The historical development of the Czochralski method. *J. Cryst. Growth* **401**, 7–24 (2014).
20. Kelso, M. V., Tubbesing, J. Z., Chen, Q. Z. & Switzer, J. A. Epitaxial electrodeposition of chiral metal surfaces on silicon(643). *J. Am. Chem. Soc.* **140**, 15812–15819 (2018).
21. Zhang, Z. et al. Identification of copper surface index by optical contrast. *Adv. Mater. Interfaces* **5**, 1800377 (2018).
22. Ateba Betanda, Y. et al. Effect of annealing atmosphere on the recrystallized texture and abnormal grain growth of Ni-5% W alloy sheets. *Adv. Eng. Mater.* **17**, 1568–1572 (2015).
23. Dunn, C. G. & Walter, J. L. Surface energies and other surface effects relating to secondary recrystallization textures in high-purity iron, zone-refined iron, and 0.6 Pct Si-Fe. *Trans. Metall. Soc. AIME* **224**, 518–533 (1962).
24. Rhead, G. E. & McLean, M. Variation of surface energy with crystallographic orientation in silver. *Acta Metall.* **12**, 401–407 (1964).
25. McLean, M. & Mykura, H. The temperature dependence of the surface energy anisotropy of platinum. *Surf. Sci.* **5**, 466–481 (1966).
26. McLean, M. Determination of the surface energy of copper as a function of crystallographic orientation and temperature. *Acta Metall.* **19**, 387–393 (1971).
27. Ma, F., Zhang, J. M. & Xu, K. W. Surface-energy-driven abnormal grain growth in Cu and Ag films. *Appl. Surf. Sci.* **242**, 55–61 (2005).
28. Mirpuri, K. et al. Texture evolution in copper film at high temperature studied in situ by electron back-scatter diffraction. *Thin Solid Films* **496**, 703–717 (2006).
29. Na, S. M. & Flatau, A. B. Global Goss grain growth and grain boundary characteristics in magnetostrictive Galfenol sheets. *Smart Mater. Struct.* **22**, 125026 (2013).
30. Birnbaum, A. J. et al. Oxygen-induced giant grain growth in Ag films. *Appl. Phys. Lett.* **111**, 163107 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Seeded growth of single-crystal high-index Cu foils

Pre-oxidation of commercial Cu foils. A polycrystalline Cu foil (25 μm thick, 99.8%, Sichuan Oriental Stars Trading Co. Ltd, #Cu-1031) was placed on a quartz substrate and loaded into a chemical vapour deposition (CVD) furnace (Tianjin Kaiheng Co. Ltd, custom designed with a heating area of length 50 cm and diameter 23 cm). The furnace was slowly heated to 150–650 °C in 10–60 min and then maintained at this temperature for typically 1–4 h to oxidize the Cu surface.

Annealing of commercial Cu foils. The pre-oxidized Cu foil was placed on a quartz substrate and loaded into an atmospheric-pressure CVD system. The system was heated to 1,020 °C in 1 h and maintained at this temperature under a gas flow of 800 standard cubic centimetres per minute (sccm) Ar and 50 sccm H_2 for 3–10 h, and was then naturally cooled to room temperature in an Ar and H_2 atmosphere. The H_2 is necessary in the annealing process and the large single-crystal growth cannot occur in a pure Ar atmosphere.

Seeded growth of single-crystal high-index Ni foils

The Ni foils (100 μm thick, 99.994%, Alfa Aesar) were first oxidized in air at 150–650 °C for 1–4 h and then annealed in a reducing atmosphere at 1,200 °C for 3–6 h. After thermal annealing, single-crystal Ni foil of size ~5 × 5 cm^2 was successfully obtained. By repeating the typical annealing procedure, several high-index single-crystal Ni foils can be produced.

Measurement of the in-plane temperature distribution of our furnace

The temperature distribution was first simulated, as shown in Extended Data Fig. 5a. The maximum temperature difference was measured by a melting experiment using Ag foils. Nine pieces of Ag foil were placed into our furnace at different positions and were then annealed at 955 °C, 960 °C and 965 °C. With increasing temperature, different parts melted (Extended Data Fig. 5b–d), and the maximum temperature difference was estimated to be ~10 °C.

Growth of graphene and hBN

Growth of graphene. The high-index single-crystal Cu foil was placed on a quartz substrate and loaded into the CVD chamber. After heating to 1,010 °C under a reducing atmosphere (500 sccm Ar, 10 sccm H_2), CH_4 (0.5 sccm) was introduced into the system for 10–30 min. Then, the system was naturally cooled under the same reducing atmosphere. The growth was carried out under atmospheric pressure.

Growth of hBN. The as-grown single-crystal Cu foil was pre-heated and re-annealed at 1,000 °C for 30 min with a gas mixture of Ar and H_2 (500 sccm Ar, 10 sccm H_2) under atmospheric pressure. A precursor of ammonia borane was nested in a quartz crucible and heated at 65 °C for sublimation, and was then carried by Ar and H_2 (250 sccm Ar, 250 sccm H_2). To visualize the individual hBN domains, the growth time was set to be 1 h. The growth was carried out under a pressure of ~200 Pa.

Characterization

The EBSD characterizations were carried out using a PHI 710 Scanning Auger Nanoprobe instrument. LEED measurements were performed using an Omicron LEED system in ultrahigh vacuum with base pressure $<3 \times 10^{-7}$ Pa. STEM experiments were performed in a FEI Titan Themis G2 300 system operated at 300 kV. SEM images were obtained using a FEI Nova NanoSEM 430 scanning electron microscope.

XRD 2θ scan measurements were conducted using a Bruker D8 Advance system with a silver target with X-ray wavelength (~0.56 Å) shorter than that of the copper target (~1.54 Å) and more suitable for high-index facet characterization with small facet spacing. XRD φ scan measurements were conducted using a PANalytical X' Pert Pro system with a copper target. Single-crystal XRD measurements were conducted using a Bruker D8 Venture system. The primitive cell of the sample was determined by single-crystal XRD, after which the facet index of the Cu foil was obtained from the reconstructed facet that was parallel to the foil surface.

STM experiments were performed with a Createc system (Germany) under ultrahigh vacuum (~1×10^{-10} torr). To remove the contamination, the Cu foils were cleaned by Ar^+ sputtering at 1 keV and annealing at 700 K for 2–5 cycles. Then, it was transferred into the liquid-nitrogen-cooled (77 K) scanner for high-resolution imaging. All the STM images were taken at the set point of 100 mV, 50 pA. Throughout the experiments, the bias voltage refers to the sample voltage with respect to the tungsten tip.

Calculation of Cu surface energy

The surface energies of various Cu facets were obtained by density functional theory (DFT) calculations as implemented in the Vienna ab initio simulation package. The exchange-correlation effect was treated using the Perdew–Burke–Ernzerhof generalized gradient approximation. The projected augmented wave method was used to describe the interaction between valence electrons and the ionic cores. To obtain the surface energy of a certain Cu facet, a thick Cu slab consisting of ≥6 atomic layers with this facet index was constructed. In addition, a thin Cu slab, which was similar to the thick one but with only half of its atomic layers, was also fabricated. After that, both the thick and the thin Cu slabs were fully optimized via the conjugated gradient method, until the force on each atom was less than 0.01 eV $Å^{-1}$. During optimization, a vacuum slab of at least 10 Å was used to avoid the periodic imaging interaction. A separation of 0.03 $Å^{-1}$ was used for the k-point mesh sampling. The energy converged to 10^{-4} eV and an energy cut-off of 400 eV for the plane-wave basis was adopted during structure optimization. The surface energy (E) of one Cu facet can be calculated as $E = (E_{thick} - 2E_{thin})/(2A)$, where E_{thick} and E_{thin} are the energy of the thick and thin Cu slabs, respectively, and A is the area of one of the two surfaces of the slab.

Calculation on grain boundary energies

We calculated the grain boundary energies to understand the abnormal grain growth in a semiquantitative way. During the annealing process, a small grain near an abnormal grain will be transformed into a part of the abnormal grain via the propagation of the grain boundary between them (Extended Data Fig. 3d). The energy change during such a grain growth process can be estimated as $\Delta E = 2S_B(\gamma_A - \gamma_B) - (\alpha L)h\gamma_{AB}$, where $2S_B \approx 2L^2$ is the surface area of grain B (considering both sides of the metal foil); L is the lateral size of grain B; αL is the estimated length of the grain boundary between them, where $\alpha \approx 1$–4 depends on the shape of grain B (the value 4 represents a single square grain inside the abnormal grain); h is the thickness of the metal foil; γ_A, γ_B and γ_{AB} are surface energies of grain A and grain B and the grain boundary energy between grains A and B, respectively. For a thermodynamically favourable annealing process, the energy change must satisfy $\Delta E = L^2[2(\gamma_A - \gamma_B) - (\alpha\gamma_{AB}h/L)] < 0$. Obviously, if $\gamma_A - \gamma_B < 0$, this requirement will be always satisfied. However, if $\gamma_A - \gamma_B > 0$, we will have $h/L > 2(\gamma_A - \gamma_B)/(\alpha\gamma_{AB})$. For our experiment $h \approx 25$ μm, and our DFT calculations using a model of a metal slab in vacuum showed that $\gamma_B \approx \gamma_{Cu(110)} \approx 0.01$ eV $Å^{-2}$ and $\gamma_{AB} \approx 0.04$ eV $Å^{-2}$. Therefore, we have $L \approx 150$ μm when γ_A is the maximum surface energy and $\alpha = 3$. Considering that the surfaces of the Cu foil must be passivated by the hydrogen during the annealing process, the surface energy difference $\gamma_B - \gamma_{Cu(110)}$ should be much smaller than our calculated value, which is based on a model in vacuum, so we expect that the annealing of a Cu foil with much large normal grains (for example, with $L > 200$ μm) is experimentally feasible.

Molecular dynamics simulation

The calculations were performed using the software package LAMMPS (large-scale atomic/molecular massively parallel simulator). The Cu–Cu

interaction was modelled with the embedded-atom method. For seeded growth case (69,695 atoms in total), a seed ($6 \times 6 \times 1.5\ nm^3$) with in-plane orientation (345) was initially placed in the middle of the polycrystalline substrate ($16 \times 16 \times 3\ nm^3$), which was divided into 8×8 sectors with distinct orientations. The system was heated from 150 K to 1,150 K at a rate of $1\ K\ ps^{-1}$, and then annealed at 1,150 K for 20 ns. Simulations were carried out using the NVT ensemble (constant number of molecules, volume and temperature), where a Nosé–Hoover thermostat was applied to fix the temperature. The velocity Verlet algorithm was used with an integration time step of 1.0 fs. Periodical boundary conditions were applied along the x and y directions.

Data availability

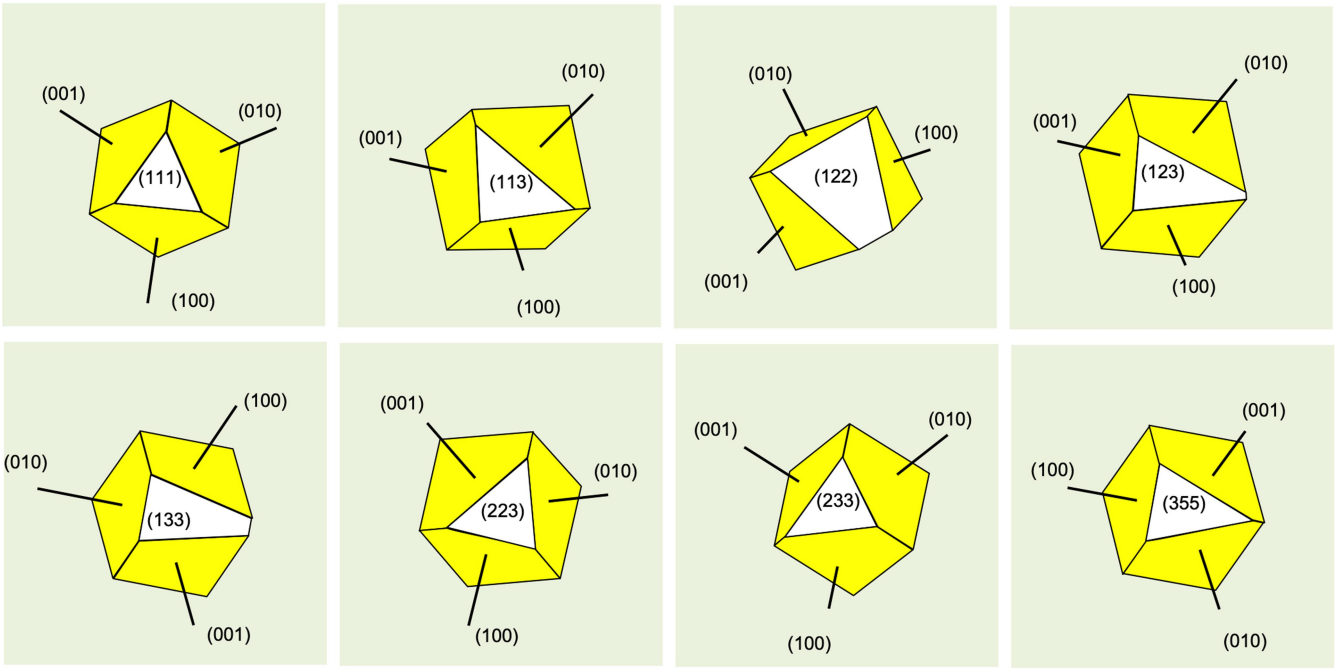
All related data generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Acknowledgements This work was supported by the National Natural Science Foundation of China (11888101, 51991340, 51991342, 21725302 and 51522201), National Key R&D Program of China (2016YFA0300903 and 2016YFA0300804), Beijing Natural Science Foundation (JQ19004), Beijing Excellent Talents Training Support (2017000026833ZK11), Beijing Municipal Science and Technology Commission (Z191100007219005), Beijing Graphene Innovation Program (Z181100004818003), Guangdong Provincial Science Fund for Distinguished Young Scholars (2020B1515020043), Bureau of Industry and Information Technology of Shenzhen (Graphene platform 201901161512), The Key R&D Program of Guangdong Province (2020B010189001, 2019B010931001, 2018B010109009 and 2018B030327001), Guangdong Innovative and Entrepreneurial Research Team Program (2016ZT06D348), the Science, Technology and Innovation Commission of Shenzhen Municipality (KYTDPT20181011104202253), National Equipment Program of China (ZDYZ2015-1), National Postdoctoral Program for Innovative Talents (BX201700014), China Postdoctoral Science Foundation (2018M630017 and 2019M660282) and the Institute for Basic Science of South Korea (IBS-R019-D1). We acknowledge the Electron Microscopy Laboratory in Peking University for the use of the electron microscope. K.L. acknowledges discussions with L. Lu.

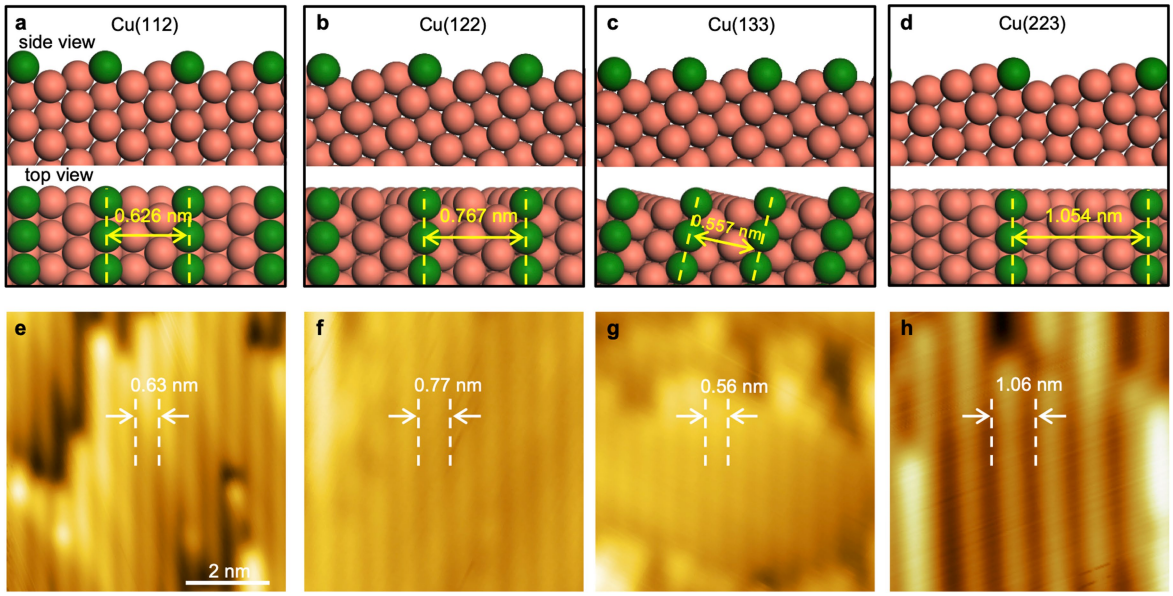
Author contributions K.L. and E.W. designed and supervised the project. K.L. conceived the experiments. E.W., F.D. and K.L. developed the growth mechanism. D.Y. organized the structural characterization. M.W., Zhibin Zhang, X.X., Zhihong Zhang, L.W., J.Q., D.Z. and N.S. fabricated the single-crystal foils. F.D., H.L., Y.D. and J.D. performed the theoretical calculations. Y.J. and S.Y. performed the STM experiments. X.B., Z.-J.W., P.G., J.S., Zhibin Zhang and R.Q. performed the SEM, TEM and STEM experiments. M.W., Zhibin Zhang, X.X., Zhihong Zhang and L.Z. performed the EBSD and LEED measurements and single-crystal XRD experiments. H.Y., Y.Y. and Zhibin Zhang performed the XRD experiments with the silver target. All of the authors discussed the results and wrote the paper.

Competing interests The authors declare no competing interests.

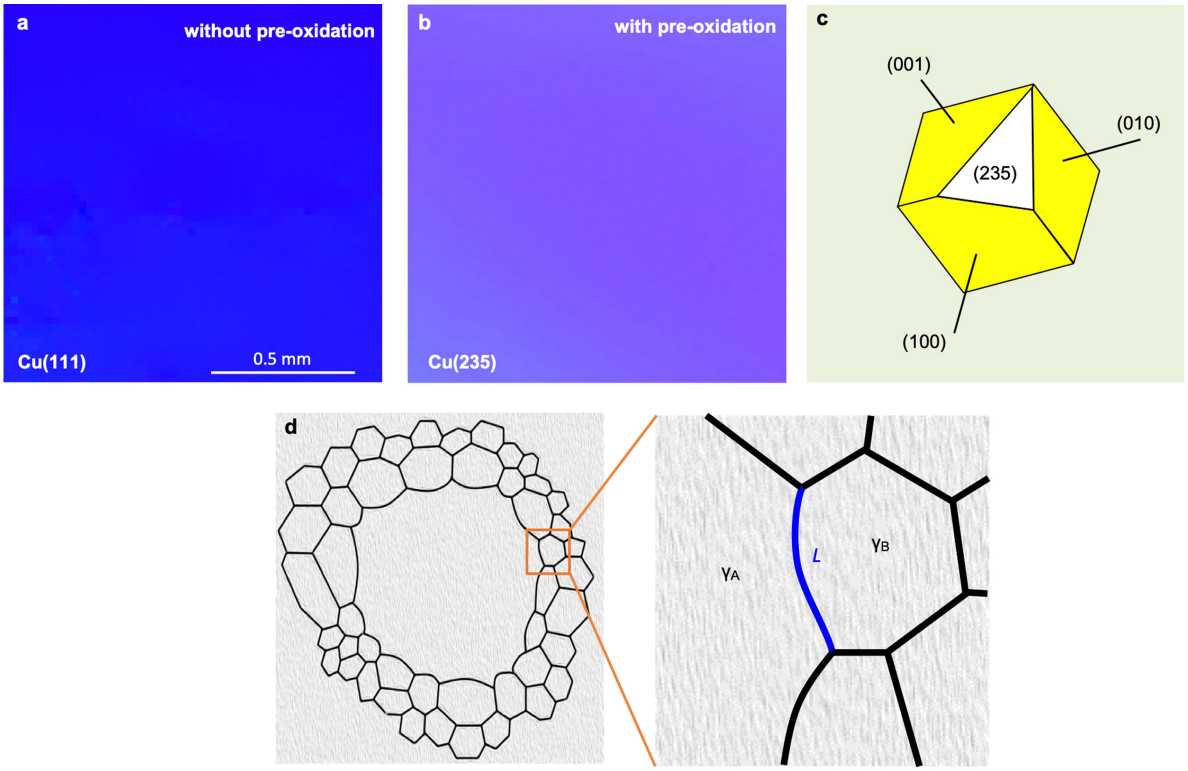
Additional information
Correspondence and requests for materials should be addressed to F.D., D.Y., E.W. or K.L.
Peer review information *Nature* thanks Jagannathan Rajagopalan, Liu Zheng and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



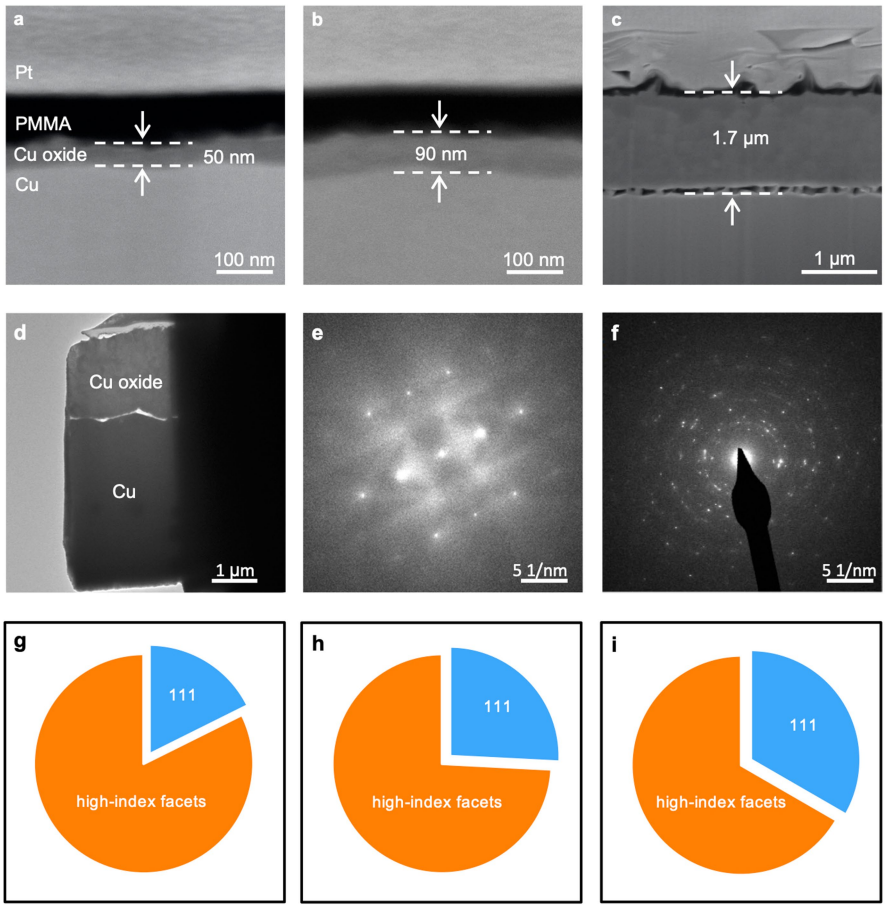
Extended Data Fig. 1 | Reconstructed single-crystal XRD images of Cu foils with various indices. Corresponding to those shown in Fig. 2. The white facets are parallel to the Cu foil surfaces.



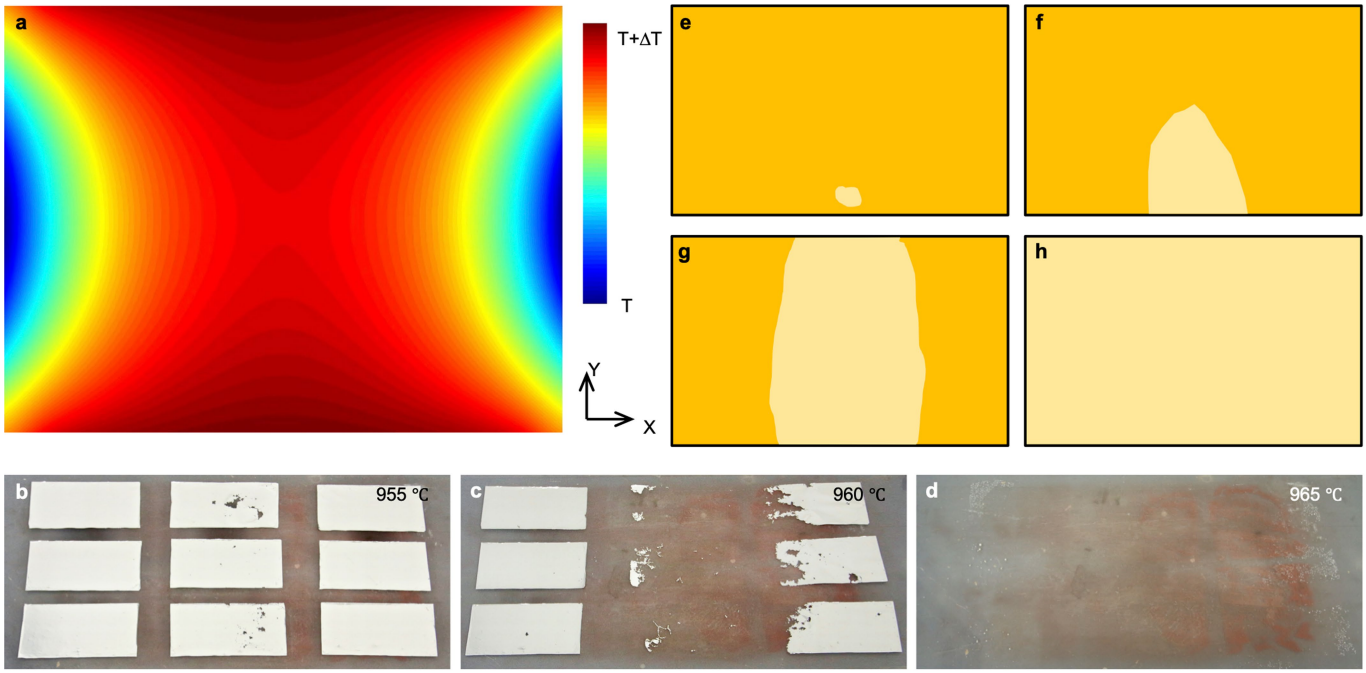
Extended Data Fig. 2 | Atomic surface structures of Cu surfaces with high-index facets. a–**d**, Atomic structures of (112) (**a**), (122) (**b**), (133) (**c**) and (223) (**d**) surfaces (top, side view; bottom, top view). **e**–**h**, Corresponding STM images of the single-crystal Cu surfaces shown in **a**–**d**. These maps are of the same size. The high-index facets feature a stripe-like structure. The measured 'stripe' spacings are consistent with calculated values from the atomic model. Because the corrugation along the Cu stripe is much smaller than that perpendicular to the stripe, owing to the strong delocalization of electrons along the stripe, the atomic resolution of Cu atoms within the stripe is not very high.



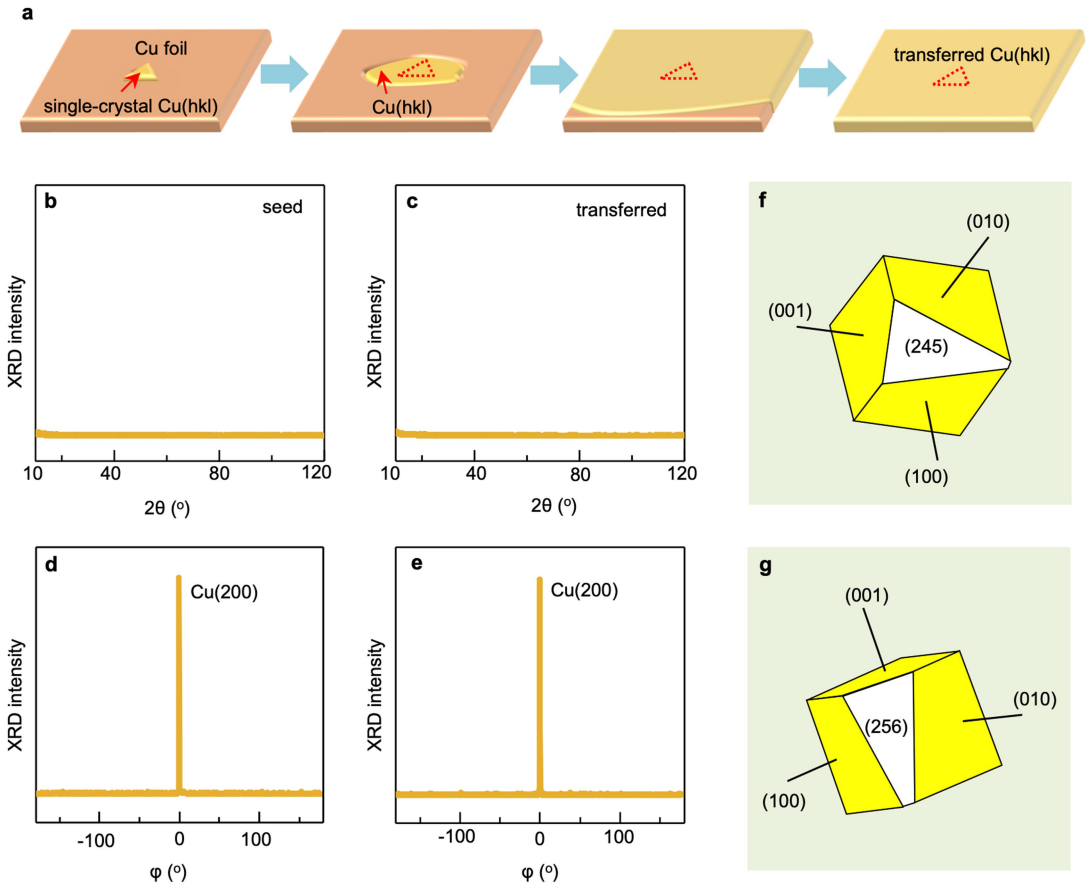
Extended Data Fig. 3 | Representative EBSD IPF maps and reconstructed single-crystal XRD image of Cu foil and the calculation model for boundary movement during abnormal grain growth. a, **b**, After annealing, the untreated terminal was transformed into Cu(111) (**a**) and the pre-oxidized part was transformed into Cu(235) (**b**). The two maps are of the same size. **c**, Reconstructed single-crystal XRD image of the Cu(235) facet. **d**, Growth of a large grain by consuming small grains around it.



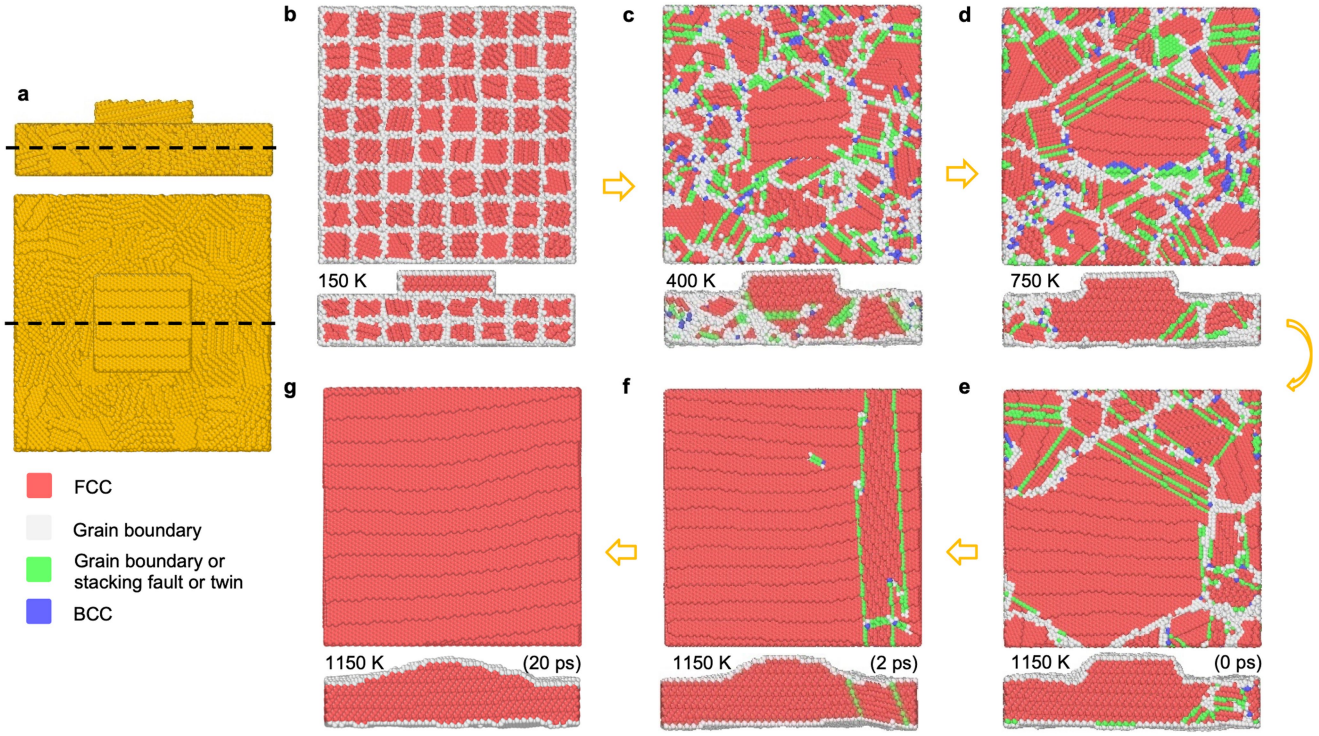
Extended Data Fig. 4 | Characterizations of the oxide layer on the Cu surface and statistics of high-index facets obtained. a–**c**, SEM images of Cu cross section after oxidation at 150 °C for 2 h (**a**), 250 °C for 2 h (**b**) and 500 °C for 1 h (**c**). The samples were prepared using the focused ion beam technique, and the polymethyl methacrylate (PMMA) and platinum (Pt) were used as protective layers to prevent damage to the sample from milling with the ion beams. **d**, Cross-sectional TEM images of Cu oxidized in air at 500 °C for 1 h. An oxidized layer was formed on the Cu surface. **e**, **f**, Selected-area electron diffraction patterns of the Cu (**e**) and the oxide layer (**f**). Unlike the single-crystal texture of the original Cu (**e**), the oxide layer has a polycrystalline structure (**f**). **g**–**i**, Statistical sector diagrams of the facet indices obtained by annealing the pre-oxidized Cu foils at conditions corresponding to **a**–**c**. The occurrence probability of high-index Cu facets was reduced from ~82% (**g**; 150 °C in air for 2 h) to ~74% (**h**; 250 °C in air for 2 h) and ~67% (**I**; 500 °C in air for 1 h), but remained very high.



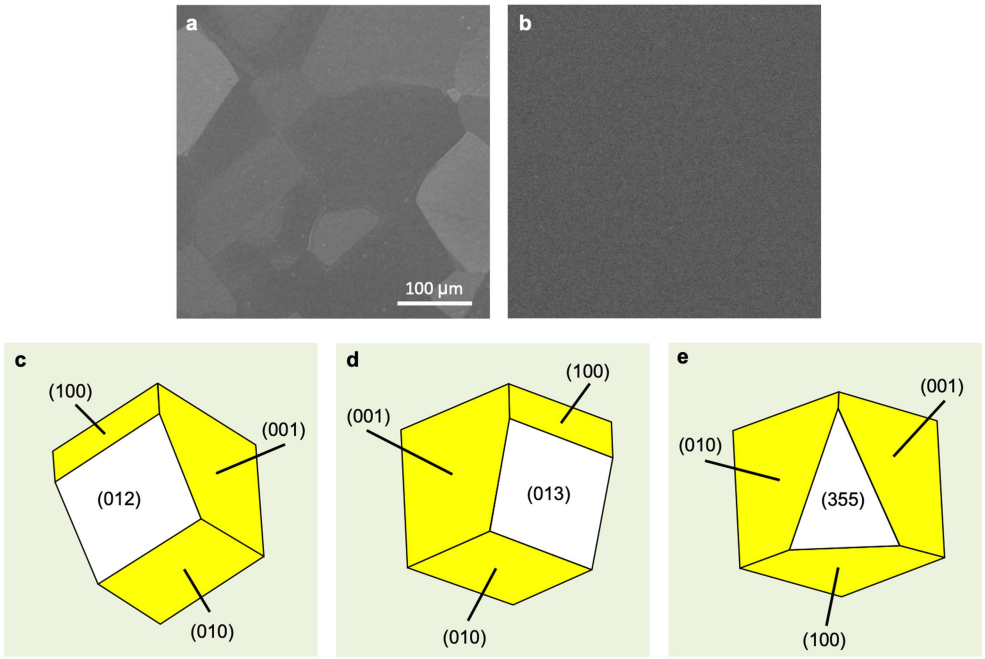
Extended Data Fig. 5 | Temperature-difference-driven single-seed abnormal grain growth. a, Simulated temperature distribution in the Cu foil. The upper and lower areas of the central part of the Cu foil have the highest temperature. **b**–**d**, Measurement of the temperature difference. Nine pieces of Ag foil were placed into our furnace and were then annealed at 955 °C (**b**), 960 °C (**c**) and 965 °C (**d**). With increasing temperature, different parts melted; the temperature difference was estimated to be ~10 °C. The image sizes for **b**–**d** are ~40 × 22 cm^2. **e**–**h**, Typical single-crystal evolution in our abnormal grain growth with a temperature difference.



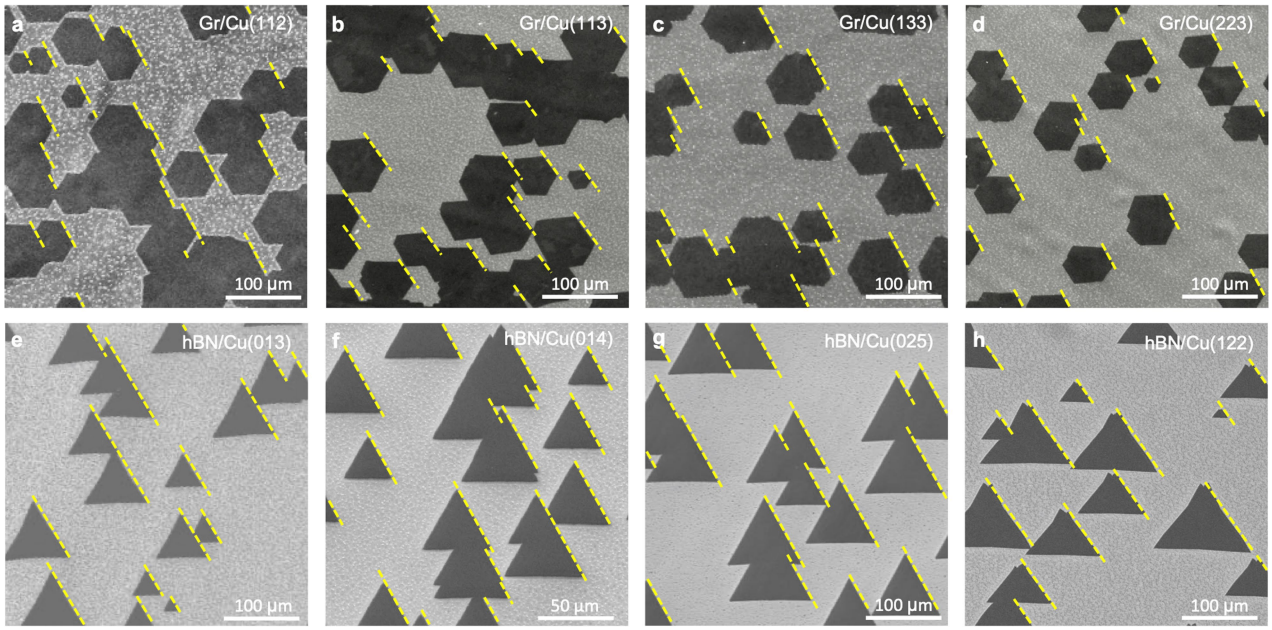
Extended Data Fig. 6 | Facet transfer and representative XRD spectra of the seed and the Cu foils with transferred facets. a, In-plane facet transfer by placing a small single-crystal Cu(*hkl*) piece on polycrystalline Cu foils. **b**–**e**, XRD 2θ (**b**) and φ (**d**) scan data for the seed, and the final Cu foil with the transferred facet (**c**, **e**). The 2θ peak for the (245) facet is out of the XRD scan range. **f**, Reconstructed single-crystal XRD image of the (245) facet, showing that the (245) seed was successfully copied and a new large-size single-crystal Cu(245) foil was produced. **g**, Reconstructed single-crystal XRD image of the Cu(256) foil.



Extended Data Fig. 7 | Simulation of facet-transfer growth. a, Simulation system. **b**–**g**, Structure evolution during the heating and preservation process. Two cross-sections are presented here, marked by dark dashed lines in **a**. **b** shows the initial configuration of the system, where the substrate is composed of 64 parts with diverse in-plane orientations. The seed is placed in the middle of the substrate surface. With increasing temperature, the part of the substrate that is in contact with the seed inherits the seed orientation first; then, the seeded grain has an advantage in size over the other grains and the orientation spreads into the entire foil (**c**–**g**). Different colours in **b**–**g** indicate distinct structures, that is, the red parts have a face-centred cubic (fcc) structure; the white parts are grain boundaries; the green parts are stacking faults, twin boundaries or grain boundaries; and the blue parts have a body-centred cubic (bcc) structure.



Extended Data Fig. 8 | Crystallographic characterization of high-index single-crystal Ni foils. a, **b**, SEM images of polycrystalline (**a**) and single-crystal (**b**) Ni foils with the same size. **c**–**e**, Reconstructed single-crystal XRD images of single-crystal Ni foils. The white facets are parallel to the Ni foil surfaces.



Extended Data Fig. 9 | Epitaxial growth of graphene and hBN on high-index single-crystal Cu foils. a–**d**, SEM images of unidirectionally aligned graphene (Gr) domains on (112) (**a**), (113) (**b**), (133) (**c**) and (223) (**d**) facets. **e**–**h**, SEM images of unidirectionally aligned hBN domains on (013) (**e**), (014) (**f**), (025) (**g**) and (122) (**h**) facets.

Extended Data Table 1 | Surface energy (in eV nm^{-2}) and facet characterization of single-crystal Cu foils

Facet index	011	012	013	014	023	025	034	111	112	113	115	122
Surface energy	9.90	9.91	9.92	9.79	9.86	9.95	9.74	8.10	9.40	9.60	9.30	9.20
2θ-scan XRD	√	√	√	√	√	√	√	√	√	√	√	√
Single-crystal XRD	√	√	√	√	√	√	√	√	√	√	√	√
Facet index	123	133	159	223	233	234	235	245	255	256	335	337
Surface energy	9.80	9.40	10.11	9.00	8.90	9.30	9.60	9.50	9.00	9.40	8.77	9.54
2θ-scan XRD	√	√	√	√	√	√	---	---	---	---	√	√
Single-crystal XRD	√	√	√	√	√	√	√	√	√	√	√	√
Facet index	344	345	346	355	356	357	359	377	455	456	577	
Surface energy	8.70	9.20	9.40	8.90	9.30	9.62	9.72	9.30	8.70	8.90	8.60	
2θ-scan XRD	---	---	---	√	---	√	√	√	---	---	√	
Single-crystal XRD	√	√	√	√	√	√	√	√	√	√	√	

'√' indicates that the facet index has been determined by this method, and '—' means that the corresponding 2θ peak for the facet is out of the XRD scan range.

Article

Photocatalytic water splitting with a quantum efficiency of almost unity

https://doi.org/10.1038/s41586-020-2278-9

Received: 26 September 2019

Accepted: 2 April 2020

Published online: 27 May 2020

Check for updates

Tsuyoshi Takata[1], Junzhe Jiang[2], Yoshihisa Sakata[2], Mamiko Nakabayashi[3], Naoya Shibata[3], Vikas Nandal[4], Kazuhiko Seki[4], Takashi Hisatomi[1] & Kazunari Domen[1,5]✉

Overall water splitting, evolving hydrogen and oxygen in a 2:1 stoichiometric ratio, using particulate photocatalysts is a potential means of achieving scalable and economically viable solar hydrogen production. To obtain high solar energy conversion efficiency, the quantum efficiency of the photocatalytic reaction must be increased over a wide range of wavelengths and semiconductors with narrow bandgaps need to be designed. However, the quantum efficiency associated with overall water splitting using existing photocatalysts is typically lower than ten per cent[1,2]. Thus, whether a particulate photocatalyst can enable a quantum efficiency of 100 per cent for the greatly endergonic water-splitting reaction remains an open question. Here we demonstrate overall water splitting at an external quantum efficiency of up to 96 per cent at wavelengths between 350 and 360 nanometres, which is equivalent to an internal quantum efficiency of almost unity, using a modified aluminium-doped strontium titanate ($SrTiO_3$:Al) photocatalyst[3,4]. By selectively photodepositing the cocatalysts Rh/Cr_2O_3 (ref. [5]) and CoOOH (refs. [3,6]) for the hydrogen and oxygen evolution reactions, respectively, on different crystal facets of the semiconductor particles using anisotropic charge transport, the hydrogen and oxygen evolution reactions could be promoted separately. This enabled multiple consecutive forward charge transfers without backward charge transfer, reaching the upper limit of quantum efficiency for overall water splitting. Our work demonstrates the feasibility of overall water splitting free from charge recombination losses and introduces an ideal cocatalyst/photocatalyst structure for efficient water splitting.

Demonstrating photocatalytic overall water splitting at an internal quantum efficiency (IQE) of 100% is an important challenge in the study of photocatalysis. Overall water splitting is a greatly uphill (Gibbs energy of +237 kJ mol^{-1}) reaction consisting of multiple electron transfer processes. To achieve a 100% IQE, the first requirement is that all photoexcited carriers must migrate to surface reaction sites before bulk recombination. In addition, two-electron injection for the hydrogen evolution reaction (HER) and four-hole injection for the oxygen evolution reaction (OER) must proceed consecutively without any backward charge transfer. However, because there are many opportunities for backward electron transfer, overall water splitting with an external quantum efficiency (EQE) greater than 50% has rarely been demonstrated, even when using ultraviolet-responsive photocatalysts[3,7–10]. Thus, it is important to establish whether a 100% IQE can be realized, by completely inhibiting backward electron transfer, and conclusively determine an effective photocatalyst structure.

$SrTiO_3$ is a suitable photocatalytic material for the assessment of this possibility. This compound is a well characterized photocatalyst with a bandgap energy of 3.2 eV (ref. [11–14]), and its EQE for overall water splitting has been improved by up to 69% over the past years using various refinements[3,4,9,14]. Here, we increased the EQE to its upper limit by constructing highly active HER and OER cocatalysts on $SrTiO_3$:Al particles site-selectively. This was accomplished by using a stepwise photodeposition method instead of an impregnation process, which results in random dispersion of the cocatalysts.

The water-splitting activity of $SrTiO_3$:Al loaded with Rh, Cr and Co species as cocatalysts via either photodeposition or conventional impregnation methods is shown in Fig. 1a. The photocatalyst loaded with Rh (0.1 wt%) and subsequently with Cr_2O_3 (0.05 wt%) via two-step photodeposition[5] evolved H_2 and O_2 at the expected stoichiometric ratio for water splitting (Fig. 1a, left). The additional photodeposition of CoOOH (0.05 wt%)[3,6] onto this sample further increased the water-splitting activity (Fig. 1a, middle). These photocatalysts, modified through sequential photodeposition, split water approximately 2 and 2.5 times faster than the sample loaded with a Rh–Cr oxide (0.1 wt% of each metal) by coimpregnation[2] (Fig. 1a, right). The highest water-splitting activity was obtained with Cr and Co loadings of 0.05 wt% each (Extended Data Fig. 1a, b). The activity was enhanced reproducibly and was maintained at 94% at least for 12.5 h (Extended Data Fig. 2). Figure 1b shows the wavelength

[1]Research Initiative for Supra-Materials, Shinshu University, Nagano, Japan. [2]Graduate School of Science and Technology for Innovation, Yamaguchi University, Ube, Japan. [3]Institute of Engineering Innovation, School of Engineering, The University of Tokyo, Tokyo, Japan. [4]Nanomaterials Research Institute, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. [5]Office of University Professors, The University of Tokyo, Tokyo, Japan. ✉e-mail: domen@shinshu-u.ac.jp

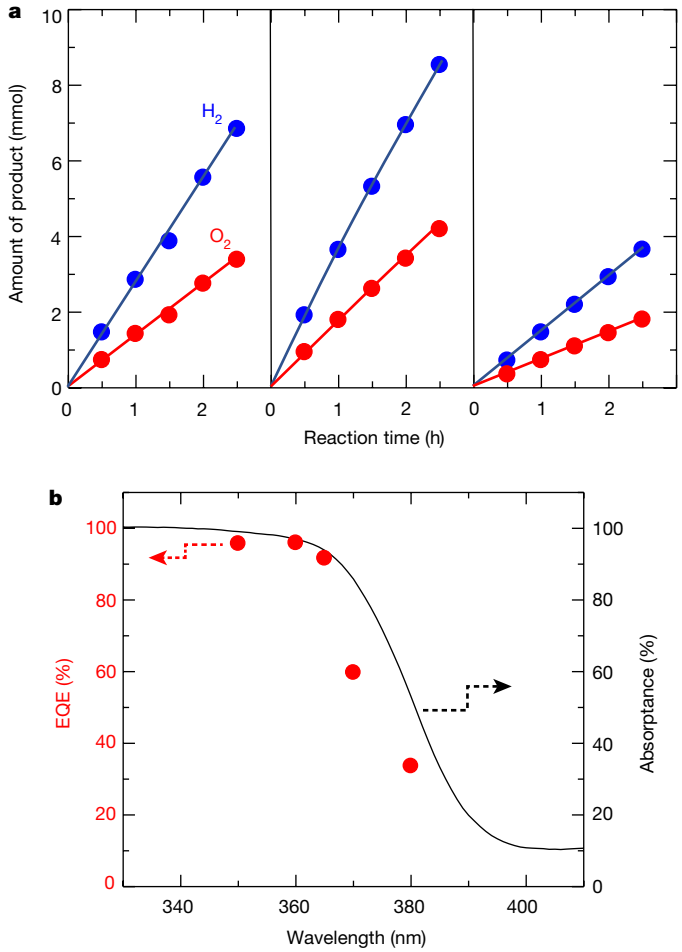


Fig. 1 | Photocatalytic water-splitting activities. a, Time course of H_2 and O_2 evolution on $SrTiO_3$:Al loaded with various cocatalysts during photoirradiation. Left, loaded with Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%) by two-step photodeposition. Middle, loaded with Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%) by three-step photodeposition. Right, loaded with Rh (0.1 wt%)–Cr (0.1 wt%) oxide by co-impregnation. Solid lines are guides for the eye. **b**, Ultraviolet–visible diffuse reflectance spectrum of bare $SrTiO_3$:Al (black solid line) and wavelength dependence of external quantum efficiency (EQE) during water splitting on Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%)-loaded $SrTiO_3$:Al (red symbols).

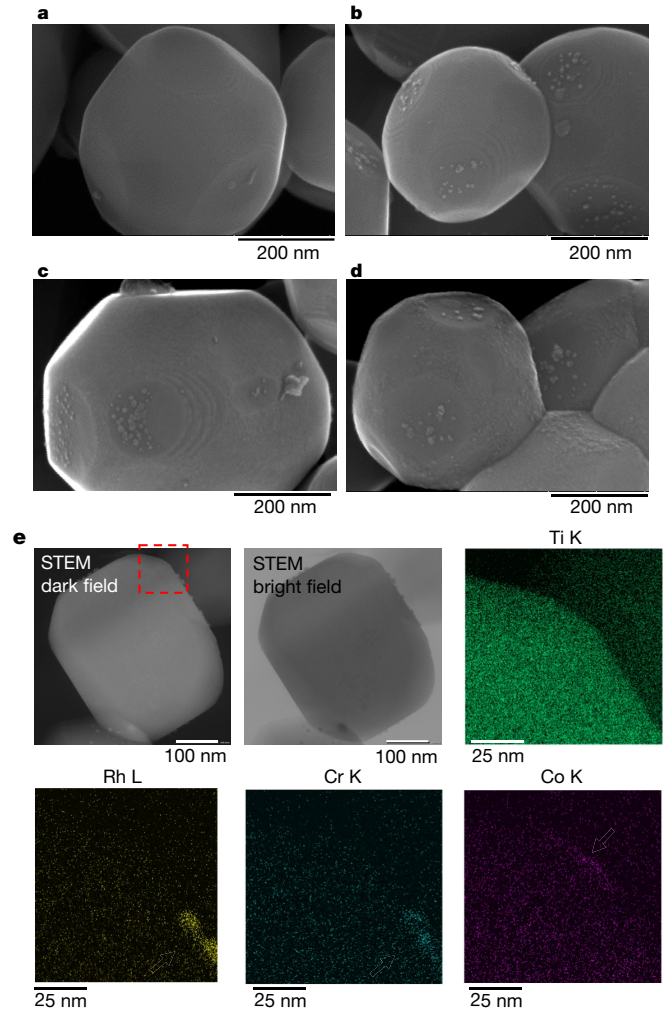


Fig. 2 | Location of cocatalysts. a–**d**, SEM images of $SrTiO_3$:Al loaded with various cocatalysts. Unloaded (**a**) and loaded with Rh (0.1 wt%) (**b**), Rh (0.1 wt%)/Cr (0.05 wt%) (**c**) and Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%) (**d**). **e**, STEM-EDS elemental mappings of $SrTiO_3$:Al loaded with Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%). EDS data were obtained from the area in the red dashed box.

dependence of the EQE during overall water splitting using the most active sample (Fig. 1a, middle) and the ultraviolet–visible diffuse reflectance spectrum of unmodified $SrTiO_3$:Al. The EQE values at 350 nm, 360 nm and 365 nm were determined to be 95.7%, 95.9% and 91.6%, respectively; to our knowledge, these are the highest values reported so far for a water-splitting photocatalyst. The EQE values at 370 nm and 380 nm were decreased to 59.7% and 33.6%, respectively, as a result of decreased light absorption and probably the lower IQE at these wavelengths. The IQE obtained on the basis of the number of absorbed photons should be close to 100% in the wavelength region of 350–360 nm considering that the EQE exceeded 95%, although it is difficult to precisely determine the number of photons absorbed by the photocatalyst owing to scattering of incident photons towards the exterior of the reactor (see Extended Data Table 1 for details). The solar-to-hydrogen (STH) efficiency under simulated sunlight illumination was 0.65% (Extended Data Fig. 3).

During the photodeposition process, photoexcited electrons and holes migrated to the surfaces of semiconductor particles and reduced or oxidized metal ions, respectively, to form deposited nanoparticles. The resulting nanoparticles acted as cocatalysts. In this case, a Rh core/Cr_2O_3 shell structure was formed during a two-step photodeposition, through the reduction of Rh^{3+} to Rh^0 and a subsequent conversion of $Cr(VI)O_4^{2-}$ to $Cr(III)_2O_3$ using photoexcited electrons[5]. We note that a shell made of Cr species is expected to be hydrated in water to form $Cr(III)O_{1.5-m}(OH)_{2m}{\cdot}xH_2O$, but herein it is denoted as Cr_2O_3 for brevity. The Rh promotes both the HER and the oxygen reduction reaction (ORR; a major backward electron transfer process), whereas the Cr_2O_3 shell inhibits only the ORR by blocking the access of evolved O_2 to the surface of the Rh core[5]. Therefore, the water-splitting rate was almost independent of the gas phase pressures (Extended Data Fig. 4). The Co^{2+} ions added as a precursor were oxidatively photodeposited as the oxyhydroxide Co(III)OOH (ref. [6]), which promotes the OER[15]. Loading none of these species or only one of the Rh, Cr, Co, Rh/Co or Cr/Co components resulted in lower or negligible photocatalytic activity (Extended Data Fig. 1c).

The microstructure of the photocatalysts modified with cocatalysts via photodeposition was investigated by electron microscopy. Figure 2a–d presents scanning electron microscopy (SEM) images of the samples at each step of the cocatalyst photodeposition. The $SrTiO_3$:Al particles were not completely cubic and various non-equivalent facets were exposed. Rh particles were deposited on specific crystal facets,

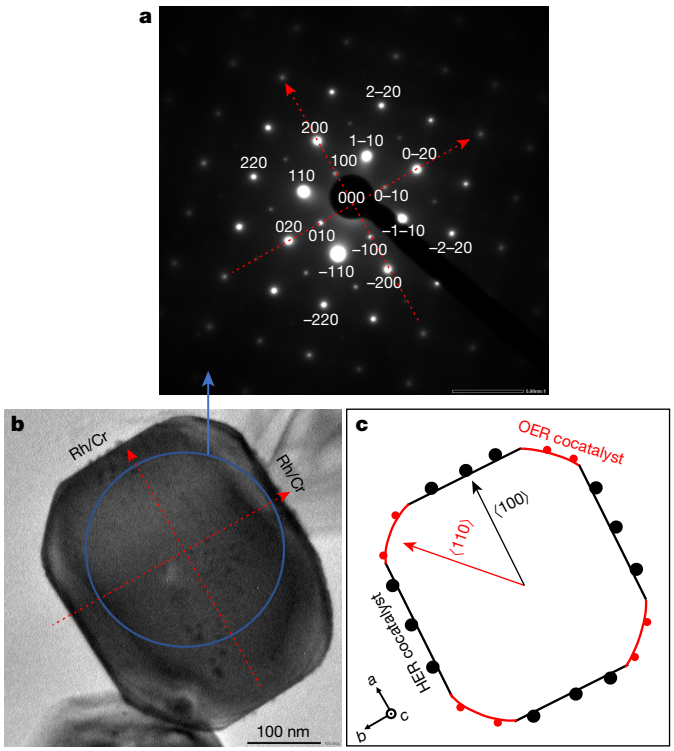


Fig. 3 | Transmission electron microscopy. a, **b**, Selected-area electron diffraction pattern obtained from $SrTiO_3$:Al loaded with Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%) (**a**) and corresponding transmission electron microscopy image of a particle (**b**). **c**, Particle morphology and crystal orientation.

and subsequent Cr_2O_3 deposition did not change the distribution of the cocatalyst particles, indicating the formation of Cr_2O_3 shells on Rh cores. Following the photodeposition of CoOOH, CoOOH nanoparticles were observed on other crystal facets, separate from the Rh/Cr_2O_3 nanoparticles. This was also confirmed by scanning transmission electron microscopy and energy-dispersive X-ray spectroscopy (STEM-EDS) analyses (Fig. 2e). Figure 3a shows the selected-area electron diffraction pattern obtained from the single-crystalline $SrTiO_3$:Al particle shown in Fig. 3b. The Rh/Cr_2O_3 cocatalyst was found to be preferentially deposited on the {100} crystal facets. The facets on which the CoOOH cocatalyst was deposited were not clearly defined because of the exposure of curved facets, but appeared to be primarily located in the ⟨110⟩ direction, as illustrated in Fig. 3c. Similar phenomena have been observed for other photocatalyst materials with anisotropic crystal habits exposing different facets[16–22]. These observations, together with the over 90% EQE values, indicate that photoexcited electrons and holes migrated to separate crystal facets, so the HER and OER subsequently occurred on these separate facets.

The observed anisotropic deposition of cocatalysts can be attributed to a charge rectification effect inside each photocatalyst particle that is induced by an internal electric field. This field, in turn, originates from the work function difference between the respective facets, just as in the p–n junction of a solar cell using the Fermi level difference. The effects of work function differences between HER and OER facets on anisotropic charge separation were simulated using a simplified two-dimensional model. The electronic energy gradient and charge distribution in a semiconductor particle with a work function difference of 0.2 eV between the {110} and {100} surfaces are plotted in Fig. 4a–e. Increasing the work function difference obviously increases the extent of anisotropic charge separation, as well as the concentration of electrons at the {100} facets and of holes at the {110} facets, as seen from Fig. 4f. A work function difference of 0.2 eV is sufficient for anisotropic charge separation (see Extended Data Fig. 5 for details). Although it is not possible at present to experimentally observe the electric field in a semiconductor particle under actual working conditions, surface dipoles resulting from unbalanced cation/anion ratios on distinct facets

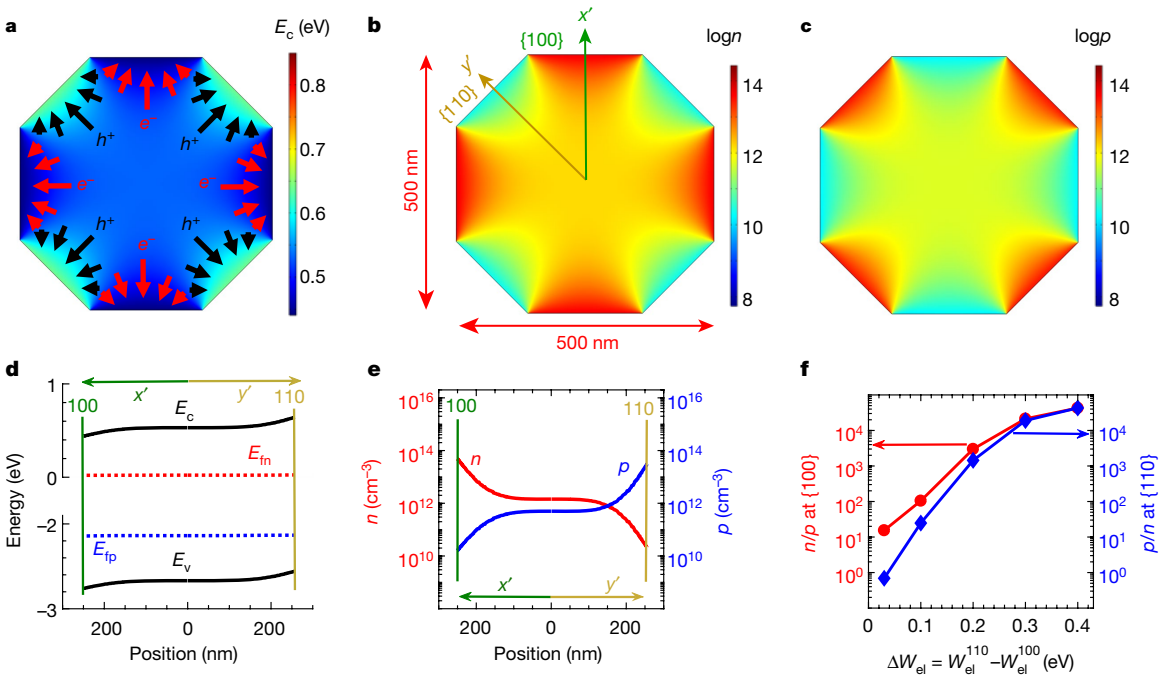


Fig. 4 | Simulations of photocarrier distributions in $SrTiO_3$:Al particles. **a**–**e**, Mapping of conduction-band energy, E_c (**a**); density of electrons (e^-), n (**b**); density of holes (h^+), p (**c**); energy band diagram (**d**); and electron and hole densities (**e**) as functions of position (x', y') with work function difference ΔW_{el} = 0.2 eV. **f**, Effect of ΔW_{el} on electron-to-hole-density ratio at the {100} and {110} facets.

may induce this work function difference even without forming junctions or composites.

The Rh/Cr_2O_3 cocatalyst was reductively photodeposited on electron-attracting {100} facets, and subsequently mediated electron transfer from the semiconductor to protons. Similarly, the CoOOH cocatalyst was oxidatively photodeposited on the hole-collecting {110} facets and neighbouring facets, and mediated hole transfer. These phenomena were central to obtaining consecutive charge transfers with minimum charge recombination for the studied cocatalyst/photocatalyst system. The concept of separating reduction and oxidation sites by facet engineering has been reported repeatedly[16–22]. However, the EQE of water splitting achieved was below 1%, and charge recombination loss was dominant. Therefore, the validity of this concept has remained controversial. This study demonstrates overall water splitting with an IQE close to unity by applying aluminium doping for defect suppression[4,14], flux treatment to improve the crystallinity[3,4], and a Cr_2O_3 shell to inhibit ORR[5], in addition to facet engineering, and thus gives a definitive answer to this problem. The selection of high-performance cocatalysts for the HER and OER is another important aspect, because prompt HER and OER suppress accumulation of photoexcited electrons and holes and the resulting recombination. This work therefore reveals a photocatalyst design that enables almost complete utilization of photoexcited electrons and holes.

In light-dependent reactions involved in photosynthesis, almost all the absorbed photons can be used to drive chemical reactions based on the functions of complex protein structures that enable prompt, one-way multistep electron transfers. At present, it is not possible to reproduce such efficient but complicated photosynthesis systems artificially. However, the particulate semiconductor system developed in this study can utilize photons at comparable quantum efficiencies during water splitting despite its simple structure. Recently, Ta_3N_5 and $Y_2Ti_2O_5S_2$ have been reported to split water into hydrogen and oxygen under visible light[23,24]. These materials absorb visible light with wavelengths of up to 600 nm and 640 nm, respectively, and the STH efficiency can reach 10% once the EQE is improved to a level similar to that of $SrTiO_3$:Al (refs. [2,5]). The suitable photocatalyst design presented here should provide impetus to the development of particulate semiconductor photocatalysts for practical solar hydrogen production from water.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2278-9.

1. Kudo, A. & Miseki, Y. Heterogeneous photocatalyst materials for water splitting. *Chem. Soc. Rev.* **38**, 253–278 (2009).
2. Chen, S., Takata, T. & Domen, K. Particulate photocatalysts for overall water splitting. *Nat. Rev. Mater.* **2**, 17050 (2017).
3. Goto, Y. et al. A particulate photocatalyst water-splitting panel for large-scale solar hydrogen production. *Joule* **2**, 509–520 (2018).
4. Ham, Y. et al. Flux-mediated doping of $SrTiO_3$ photocatalysts for efficient overall water splitting. *J. Mater. Chem. A* **4**, 3027–3033 (2016).
5. Maeda, K. & Domen, K. Photocatalytic overall water splitting: recent progress and future challenges. *J. Phys. Chem. Lett.* **1**, 2655–2661 (2010).
6. Lyu, H. et al. An Al-doped $SrTiO_3$ photocatalyst maintaining sunlight-driven overall water splitting activity over 1,000 h of constant illumination. *Chem. Sci.* **10**, 3196–3201 (2019).
7. Sakata, Y., Hayashi, T., Yasunaga, R., Yanaga, N. & Imamura, H. Remarkably high apparent quantum yield of the overall photocatalytic H_2O splitting achieved by utilizing Zn ion added Ga_2O_3 prepared using dilute $CaCl_2$. *Chem. Commun.* **51**, 12935–12938 (2015).
8. Kato, H., Asakura, K. & Kudo, A. Highly efficient water splitting into H_2 and O_2 over lanthanum-doped $NaTaO_3$ photocatalysts with high crystallinity and surface nanostructure. *J. Am. Chem. Soc.* **125**, 3082–3089 (2003).
9. Chiang, T. H. et al. Efficient photocatalytic water splitting using Al-doped $SrTiO_3$ coloaded with molybdenum oxide and rhodium-chromium oxide. *ACS Catal.* **8**, 2782–2788 (2018).
10. Li, Y. et al. Photocatalytic water splitting by N-TiO_2 on MgO(111) with exceptional quantum efficiency at elevated temperatures. *Nat. Commun.* **10**, 4421 (2019).
11. Wagner, F. T. & Somorojai, G. A. Photocatalytic and photoelectrochemical hydrogen production on strontium titanate single crystals. *J. Am. Chem. Soc.* **102**, 5494–5502 (1980).
12. Domen, K. et al. Photocatalytic decomposition of water vapour on an NiO-$SrTiO_3$ catalyst. *J. Chem. Soc. Chem. Commun*. 543–544 (1980).
13. Lehn, J. M., Sauvage, J. P., Zlessel, R. & Hilaire, L. Water photolysis by UV irradiation of rhodium loaded strontium titanate catalysts. Relation between catalytic activity and nature of the deposit from combined photolysis and ESCA studies. *Isr. J. Chem.* **22**, 168–172 (1982).
14. Takata, T. & Domen, K. Defect engineering of photocatalyst by doping of aliovalent metal cations for efficient water splitting. *J. Phys. Chem. C* **113**, 19386–19388 (2009).
15. Kanan, M. W. & Nocera, D. G. In situ formation of an oxygen-evolving catalyst in neutral water containing phosphate and Co^{2+}. *Science* **321**, 1072–1075 (2008).
16. Ohno, T., Sarukawa, K. & Matsumura, M. Crystal faces of rutile and anatase TiO_2 particles and their roles in photocatalytic reactions. *New J. Chem.* **26**, 1167–1170 (2002).
17. Liu, G., Yu, J. C., Lu, G. Q. & Cheng, H. M. Crystal facet engineering of semiconductor photocatalysts: motivations, advances and unique properties. *Chem. Commun.* **47**, 6763–6783 (2011).
18. Xie, Y. P., Liu, G., Yina, L. & Cheng, H. M. Crystal facet-dependent photocatalytic oxidation and reduction reactivity of monoclinic WO_3 for solar energy conversion. *J. Mater. Chem.* **22**, 6746–6751 (2012).
19. Zhen, C., Yu, J. C., Liu, G. & Cheng, H. M. Selective deposition of redox co-catalyst(s) to improve the photocatalytic activity of single-domain ferroelectric $PbTiO_3$ nanoplates. *Chem. Commun.* **50**, 10416–10419 (2014).
20. Li, R. et al. Spatial separation of photogenerated electrons and holes among {010} and {110} crystal facets of $BiVO_4$. *Nat. Commun.* **4**, 1432 (2013).
21. Mu, L. et al. Enhancing charge separation on high symmetry $SrTiO_3$ exposed with anisotropic facets for photocatalytic water splitting. *Energy Environ. Sci.* **9**, 2463–2469 (2016).
22. Luo, Y. et al. Construction of spatial charge separation facets on $BaTaO_2N$ crystals by flux growth approach for visible-light-driven H_2 production. *ACS Appl. Mater. Interfaces* **11**, 22264–22271 (2019).
23. Wang, Z. et al. Overall water splitting by Ta_3N_5 nanorod single crystals grown on the edges of $KTaO_3$ particles. *Nat. Catal.* **1**, 756–763 (2018).
24. Wang, Q. et al. Oxysulfide photocatalyst for visible-light-driven overall water splitting. *Nat. Mater.* **18**, 827–832 (2019).

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Synthesis of Al-doped $SrTiO_3$

$SrTiO_3$ doped with Al ($SrTiO_3$:Al) was synthesized using molten-salt mediation, according to a previously reported method[3]. Briefly, $SrCl_2$ (Kanto Chemical Co.), Al_2O_3 (Merck) and $SrTiO_3$ (Wako Pure Chemical Industries) were mixed by grinding in an agate mortar in a 10:0.02:1 molar ratio. The mixture was subsequently heated in an alumina crucible at 1,423 K for 10 h in air and then allowed to cool to room temperature. The product was then separated from residual $SrCl_2$ and Al_2O_3 and related compounds by washing with distilled water three times, at which point the supernatant solution was neutral. The amount of Al incorporated in the $SrTiO_3$ via this process was slightly lower than one atom per cent with respect to Ti (ref. [3]).

Cocatalyst loading

Photodeposition was performed in situ by first dispersing $SrTiO_3$:Al (0.1 g) in 100 ml of distilled water in the same vessel employed for the photocatalytic reactions, using a brief sonication. Aqueous solutions of $RhCl_3{\cdot}6H_2O$ (Wako Pure Chemical Industries), K_2CrO_4 (Kanto Chemical Co.) and $Co(NO_3)_2{\cdot}6H_2O$ (Kanto Chemical Co.) were prepared. We note that the aqueous solution of $RhCl_3{\cdot}6H_2O$ should be prepared freshly before use. The concentration of the metal species (Rh, Cr and Co) was 2 mg ml^{-1}. Calculated amounts of these aqueous solutions were added into the reaction solution for successive photodeposition. To deposit 0.1 wt% Rh, 0.05 wt% Cr and 0.05 wt% Co on 0.1 g of $SrTiO_3$:Al, the $RhCl_3$ aqueous solution (50 μl) was first added to the reaction suspension with magnetic stirring, and the resulting mixture was irradiated with a Xe lamp (300 W, full arc; details in Extended Data Fig. 6) for 10 min. Subsequently, the K_2CrO_4 aqueous solution (25 μl) was added to the suspension with additional irradiation for 5 min, followed by the addition of a specific quantity of $Co(NO_3)_2$ solution and another 5-min irradiation. We note that photodeposition of these components was performed in a sequence without exchanging the solvent. A cocatalyst based on a mixture of Rh–Cr oxides was also deposited on the main catalyst using a previously reported impregnation method[2]. In this process, $SrTiO_3$:Al was dispersed in a small amount of distilled water containing specific amounts of Na_3RhCl_6 (Mitsuwa Chemicals Co.) and $Cr(NO_3)_3$ (Kanto Chemical Co.), after which the mixture was heated on a hot water bath until dry, followed by heating in air for 1 h at 623 K. Regardless of the chemical state of the deposited cocatalyst, the loading amount of each component was calculated assuming a metallic state for simplicity.

Characterization

The crystal structure of the $SrTiO_3$:Al was evaluated by X-ray diffraction (XRD; Cu Kα radiation, Miniflex 300, Rigaku Co.; see Extended Data Fig. 7). The optical absorption spectrum of the $SrTiO_3$:Al (Fig. 1b) was determined using a ultraviolet–visible spectrometer equipped with an integrating sphere (V-670, JASCO), employing a spectralon block as a reference to adjust the 100% reflectance level and set the reflectance of the sample holder to 0%. Absorptance was given by 1 − reflectance. The microstructure of the photodeposited samples was analysed by field-emission SEM (SU-8020, Hitachi High-Technologies Co.) and by field-emission (scanning) transmission electron microscopy (JEM-2800F, JEOL) in conjunction with EDS using an X-MAX 100TLE SDD detector (Oxford Instruments). Selected-area electron diffraction patterns were indexed using a ReciPro diffraction simulator.

Photocatalytic reactions

Photocatalytic reactions were carried out in an overhead-irradiation-type glass vessel connected to a closed gas circulation system. Prior to each reaction, all air was evacuated from the reaction system and filled with Ar (about 1 kPa unless otherwise noted). The suspension was subsequently irradiated using a Xe lamp (300 W, full arc). Evolved gases accumulated in the closed gas circulation system were analysed by gas chromatography (GC-8A, Shimadzu Co., thermal conductivity detector, Ar carrier gas, molecular sieve 5 Å column). The STH efficiency was measured under simulated sunlight irradiation (AM1.5G, 9 cm^2 illuminated area, solar simulator HAL-320, Asahi Spectra Co.). The STH efficiency was determined according to the following equation

$$\mathrm{STH}(\%) = \frac{r_{H_2} \times \Delta G_{H_2O}}{I \times S} \times 100 \quad (1)$$

Here, r_{H_2}, ΔG_{H_2O}, I and S represent the H_2 evolution rate, the reaction Gibbs energy of water splitting, the light energy flux and the irradiation area, respectively.

Measurement of quantum efficiencies

The EQE values were determined according to the following equation

$$\mathrm{EQE}(\%) = 2 \times \frac{N(H_2)}{N(\mathrm{photons})} \times 100 \quad (2)$$

where $N(H_2)$ and N(photons) denote the number of H_2 molecules produced and the number of photons reaching the surface of the reaction solution, respectively. The H_2 evolution rate was measured in the same reaction system as the other photocatalytic reactions. To determine the wavelength dependence of the EQE, samples were irradiated with monochromatic light generated by a Xe lamp and sent through bandpass filters with central wavelengths of 350 nm, 360 nm, 365 nm, 370 nm or 380 nm (Edmund Optics). The full-width at half-maximum of each of these bandpass filters was approximately 10 nm. The dependence of the EQE on the irradiance was established using monochromatic light at 365 nm passed through a bandpass filter and neutral density filters (Edmund Optics); see Extended Data Table 1 for details. The numbers of photons were counted using photon-to-current conversion with a Si photodiode and a multimeter in the device shown in Extended Data Fig. 8. The photon flux, I, was calculated from the photocurrent density generated by a Si photodiode, Photocurrent(Si), measured using the device shown in Extended Data Fig. 8a–c, according to the equation

$$I = \mathrm{Photocurrent(Si)} \times \frac{N_A}{F} \times \mathrm{QE(Si)} \quad (3)$$

where N_A, F and QE(Si) represent the Avogadro constant, the Faraday constant and the quantum efficiency of the Si photodiode, respectively. A cylindrical reactor with an inner diameter of 65 mm with an optically flat window made of quartz and a Xe lamp (300 W) were employed for water splitting. Because the Xe lamp emits divergent light, I is not uniform over the entire light acceptance area. To accurately measure N(photons), a circular mask with an inner diameter of 30 mm was inserted to narrow the distribution of the light intensity. Then, I was measured at various positions by sliding the position of a Si photodiode with an interval of 2 mm, with the level of the Si photodiode maintained at the surface of the reaction solution. This gives $I(r)$, I as a function of the distance from the centre, r. N(photons) can be then calculated by numerically integrating $I(r)$ over the entire light acceptance area of the reactor

$$N(\mathrm{photons}) = I(0) \times \pi r_1^2 + \sum_{k=1}^{18} I\left(\frac{r_{k+1} + r_k}{2}\right) \times \pi(r_{k+1}^2 - r_k^2) \quad (4)$$
$$r_k = 2k - 1$$

where k is a non-negative integer. $I(r)$ was taken as the average of the intensities measured at two points symmetrical to the centre of the circle (see Extended Data Fig. 8d).

Electrical simulations of $SrTiO_3$:Al particulate systems

$SrTiO_3$:Al with a perovskite crystal structure was modelled using an octagonal geometry based on SEM images of $SrTiO_3$:Al nanoparticles,

in which the OER and HER proceed at diagonal and non-diagonal edges, respectively. In this model, we defined the x' and y' directions relative to the non-diagonal {100} and diagonal {110} edges, respectively. The $SrTiO_3$:Al material parameters required for electrical simulations were obtained from the literature and are provided in Extended Data Table 2[25–30]. Charge carrier recombination was simulated based on defect-assisted Shockley–Reed–Hall recombination[31], with the defect states located in the middle of the bandgap of $SrTiO_3$:Al. The trapping times or lifetimes of the electrons and holes were fixed so that the diffusion lengths of the charge carriers exceeded the particle size, given that the experimentally observed IQE values were nearly 100%. The illumination conditions were reproduced by providing a uniform generation rate across the particle dimensions, because the random orientation of crystal facets and multiple scattering in the suspension hindered the applicability of the Beer–Lambert law. However, this assumption does not affect the band diagram and hence the description of anisotropic charge transport qualitatively. The boundary conditions were provided by two distinct pseudo Schottky contacts with work functions W_{el}^{100} and W_{el}^{110} at the non-diagonal and diagonal edges, respectively. The W_{el}^{100} value was adjusted to equal the midpoint between the redox potentials for H_2 and O_2 evolution and further tuned in accordance with the pH level of the electrolyte (pH = 7 for water). On this basis, we obtained W_{el}^{100} = 4.64 eV relative to the vacuum energy level at a pH of 7 as per the Nernst equation. However, it is known that charging or dipole effects can produce differences in the work functions for the {100} and {110} edges. To include such effects, we fixed the value of W_{el}^{100} and varied W_{el}^{110} from 4.64 eV to 5.05 eV. The extraction rates for electrons (for the HER) and holes (for the OER) were non-zero for the non-diagonal and diagonal edges, respectively. Using COMSOL Multiphysics, the model was discretized by physics-controlled meshing that produced an extremely fine mesh near the edges and a relatively fine mesh throughout the particle bulk. The COMSOL Multiphysics Semiconductor Module was used to perform electrical simulations that solved the Poisson, drift–diffusion and continuity equations self-consistently for electrons and holes at each discretized node. Energy band diagrams and electron and hole densities were plotted in one dimension along the x' and y' directions to show their respective variations from the {100} to {110} edges.

Extended Data Figure 5 shows the effect of the work function difference $\Delta W_{el} = W_{el}^{110} - W_{el}^{100}$ on the energy band diagram and charge carrier density under illumination. The accumulation and depletion regions that appear near the crystal facets are primarily determined by the relative positions of the work functions of the pseudo metal contacts (W_{el}^{100}, W_{el}^{110}) and of the semiconductor (W_s). As an example, in the case of W_{el}^{100}, $W_{el}^{110} < W_s$, the semiconductor accumulates electrons from the pseudo metal contacts at the {100} and {110} facets, resulting in a flat energy band diagram for ΔW_{el} = 0.03 eV (with the given particle dimensions). This work function difference does not induce selective transport of charge carriers towards different crystal facets. However, $W_{el}^{100} < W_s < W_{el}^{110}$ produces accumulation regions near the {100} facets and depletion regions in the vicinity of the {110} facets when $\Delta W_{el} > 0.1$ eV. The formation of these distinct regions produces downward and upward (asymmetric) band bending around the {100} and {110} facets, respectively, which leads to anisotropic charge transport. The extent of asymmetric band bending, and hence anisotropic charge transport, in the vicinity of the {100} and {110} facets increases with increasing ΔW_{el}. Furthermore, for $\Delta W_{el} < 0.4$ eV, we find that the hole density is less than the electron density in the bulk of the particulate system owing to substantial electron injection from the pseudo metal contacts associated with the {100} facets. By contrast, for ΔW_{el} = 0.4 eV, electron injection into the semiconductor bulk is suppressed by increases in the depletion regions originating from W_{el}^{110} around the {110} facets. As a result, the bulk electron and hole densities become almost equal.

Data availability

Supporting data are available at the Shinshu University Institutional Repository at http://hdl.handle.net/10091/00021822.

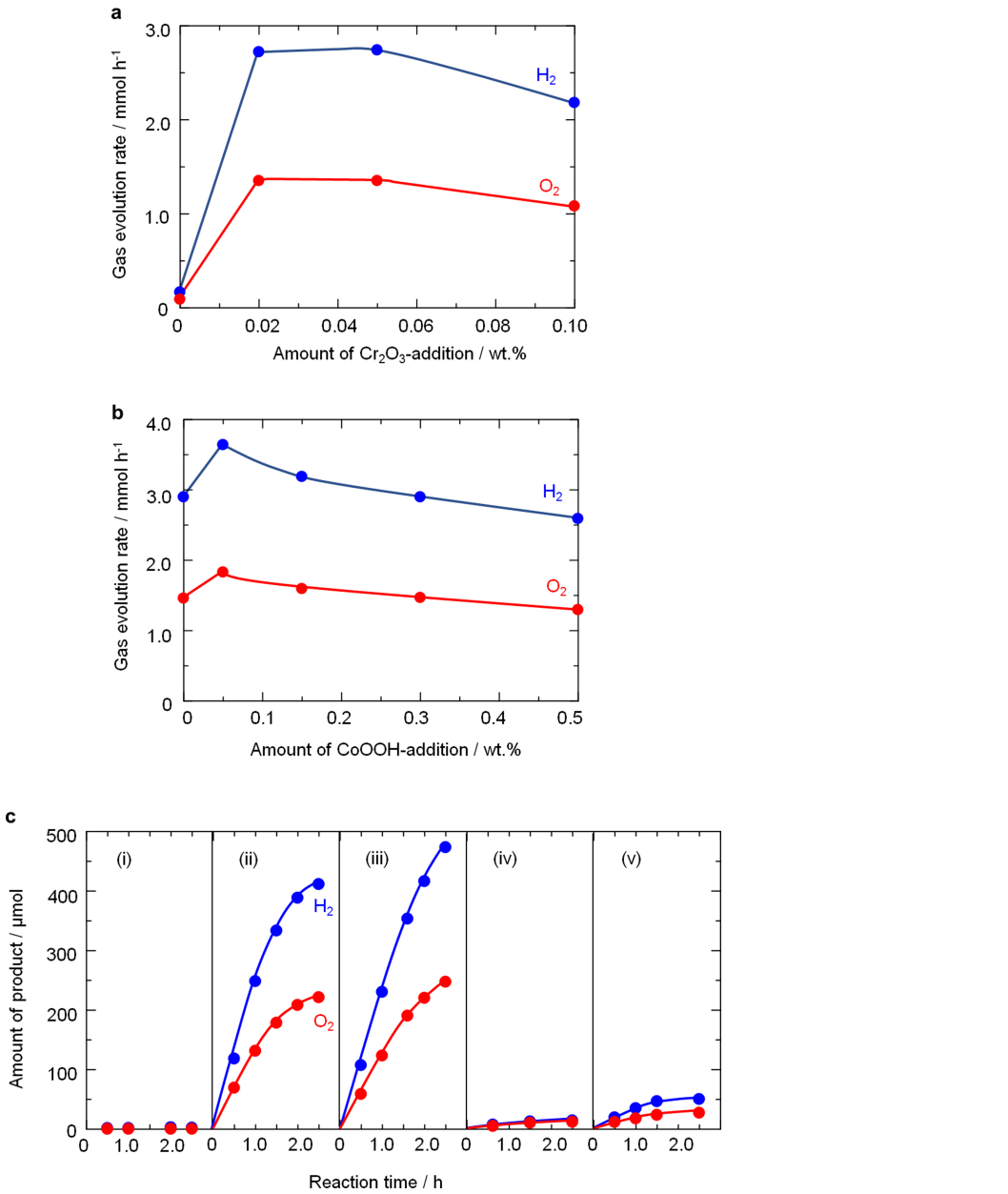
25. Frederikse, H., Thurber, W. & Hosler, W. Electronic transport in $SrTiO_3$. *Phys. Rev.* **134**, A442–A445 (1964).
26. Weaver, H. Dielectric properties of single crystals of $SrTiO_3$ at room temperatures. *J. Phys. Chem. Solids* **11**, 275–277 (1959).
27. Zhao, Z. et al. Electronic structure basis for enhanced overall water splitting photocatalysis with aluminum doped $SrTiO_3$ in natural sunlight. *Energy Environ. Sci.* **12**, 1385–1395 (2019).
28. Konta, R., Ishii, T., Kato, H. & Kudo, A. Photocatalytic activities of noble metal ion doped $SrTiO_3$ under visible light irradiation. *J. Phys. Chem. B* **108**, 8992–8995 (2004).
29. Cohen, M. I. & Blunt, R. F. Optical properties of $SrTiO_3$ in the region of absorption edge. *Phys. Rev.* **168**, 929–933 (1968).
30. Zollner, S. et al. Optical properties of bulk and thin film $SrTiO_3$ on Si and Pt. *J. Vac. Sci. Technol. B* **18**, 2242–2254 (2000).
31. Sze, S. M. *Physics of Semiconductor Devices* (John Wiley & Sons, 1981).

Acknowledgements This work was primarily supported by the Artificial Photosynthesis Project of the New Energy and Industrial Technology Development Organization (NEDO) and partly supported by JSPS KAKENHI grant number JP19K05669, JP16K06862. A part of this work was conducted at the Advanced Characterization Nanotechnology Platform of the University of Tokyo, supported by "Nanotechnology Platform" of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan (grant number JPMXP09A-19-UT-0023). We are grateful to M. Yamaguchi and Y. Kuromiya at the University of Tokyo for the preparation and evaluation of photocatalysts.

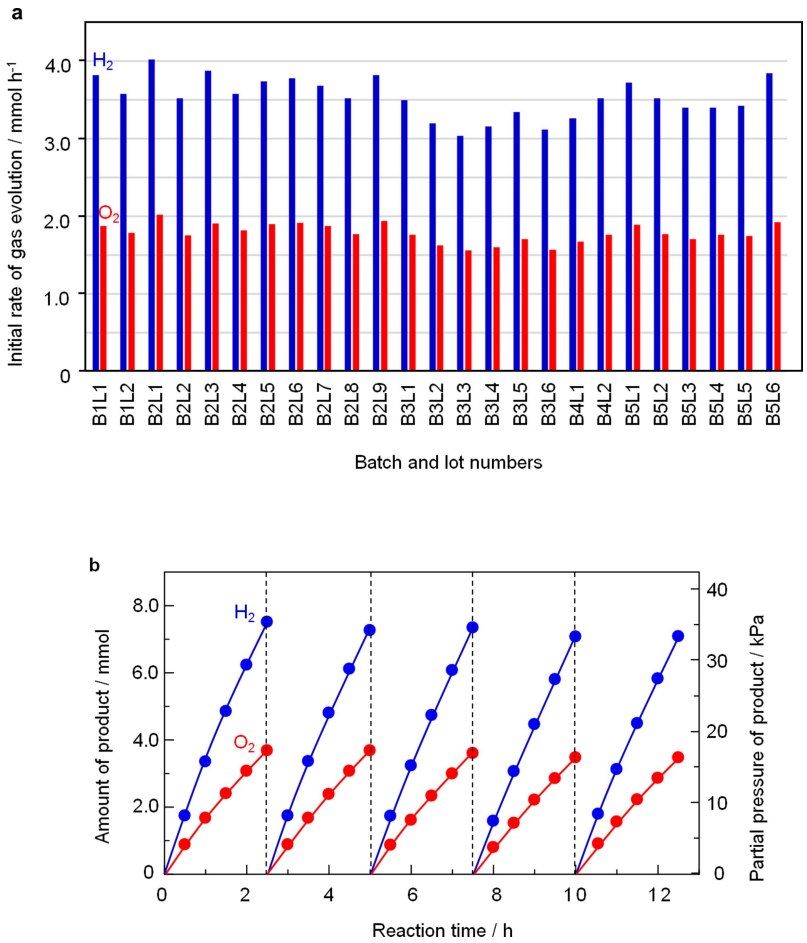
Author contributions T.T. conceived the photocatalyst design and performed the photocatalytic reactions. T.H. synthesized Al-doped $SrTiO_3$. J.J. and Y.S. performed quantum efficiency measurements. M.N. and N.S. performed electron microscopy measurements. V.N. and K.S. performed electrical simulations. K.D. supervised the project. T.T., T.H. and K.D. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

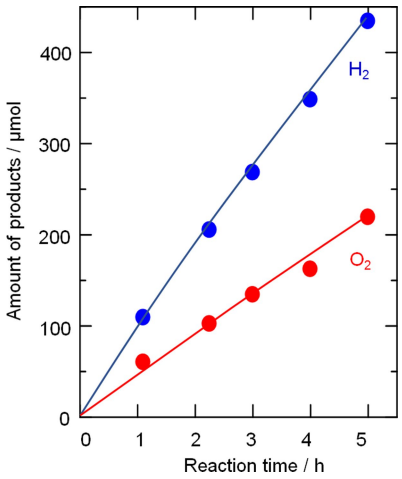
Additional information
Correspondence and requests for materials should be addressed to K.D.
Peer review information *Nature* thanks David Tilley, Martijn Zwijnenburg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



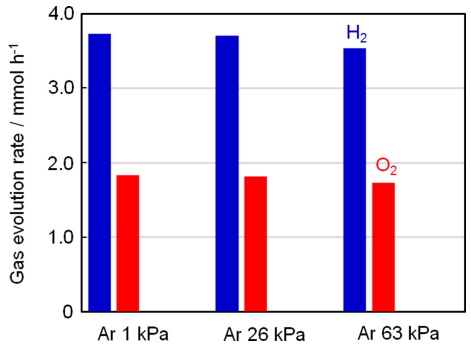
Extended Data Fig. 1 | Dependence of water-splitting activities on the amounts and components of cocatalysts. a, **b**, Dependence of water-splitting activity on the amount of Cr_2O_3 added to Rh (0.1 wt%)-loaded $SrTiO_3$:Al (**a**) and CoOOH added to Rh (0.1 wt%)/Cr_2O_3(0.05 wt%)-loaded $SrTiO_3$:Al (**b**). **c**, Time course of H_2 and O_2 evolution over $SrTiO_3$:Al unloaded (i) and loaded with Rh (0.1 wt%) (ii), Rh (0.1 wt%)/CoOOH (0.05 wt%) (iii), CoOOH (0.05 wt%) (iv) and CoOOH (0.05 wt%)/Cr_2O_3 (0.05 wt%) (v).



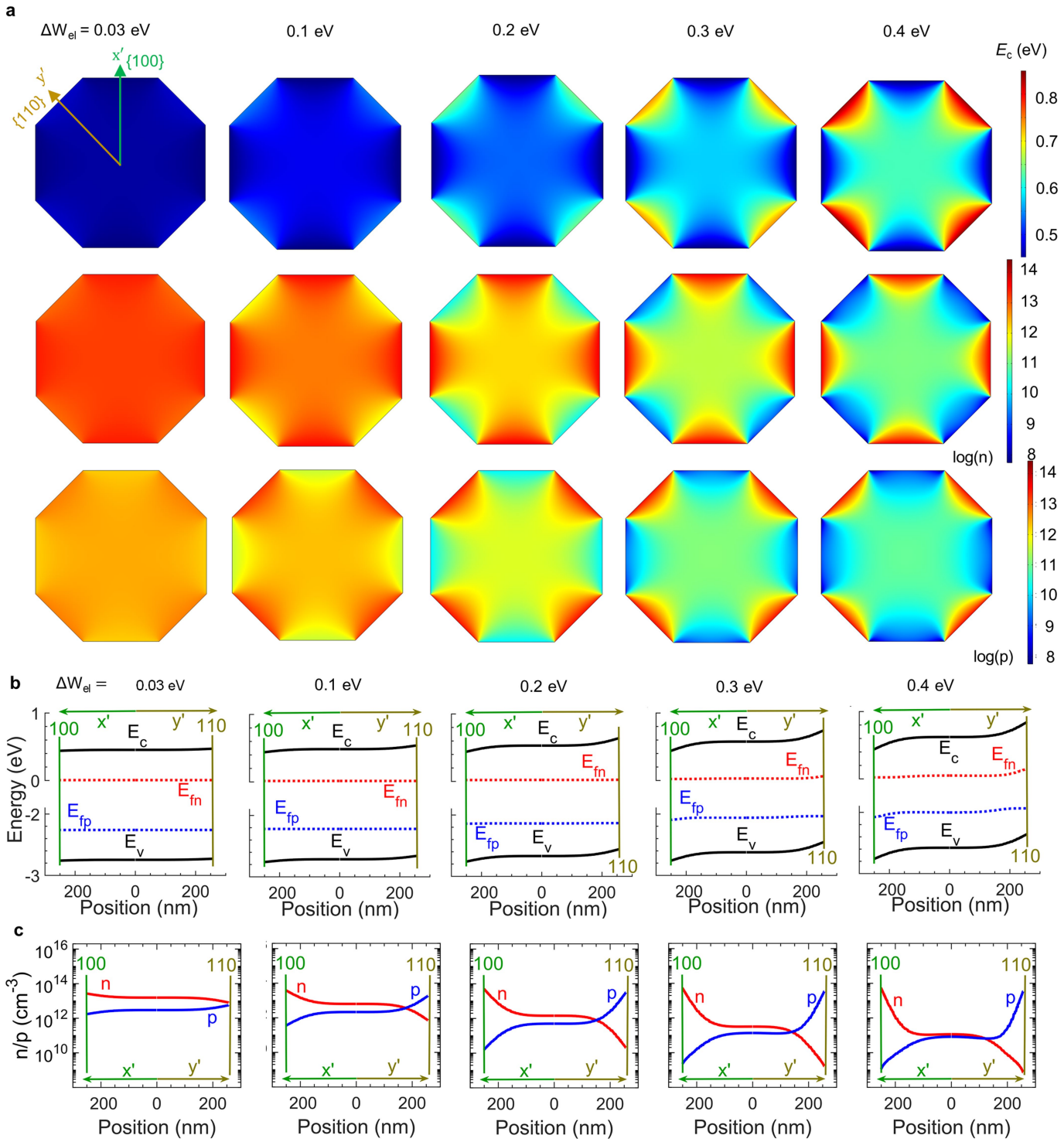
Extended Data Fig. 2 | Reproducibility and stability of water-splitting activity of $SrTiO_3$:Al loaded with Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%). a, Variation of water-splitting activity of $SrTiO_3$:Al loaded with Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%) (25 samples) from different batches (B) and lots (L). Batches and lots represent $SrTiO_3$:Al samples and the cocatalysts synthesized and loaded at different timings, respectively. The averages (μ) of the H_2 and O_2 evolution rates were 3.54 mmol h^{-1} and 1.78 mmol h^{-1} and their standard deviations (σ) were 0.26 mmol h^{-1} and 0.12 mmol h^{-1}, respectively. All the gas evolution rates were within $\mu \pm 2\sigma$. **b**, Stability test of $SrTiO_3$:Al loaded with Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%) during photocatalytic water splitting. The produced H_2 and O_2 accumulated in the reaction system were evacuated every 2.5 h. The inner wall of the reactor window was cleaned before every evacuation.



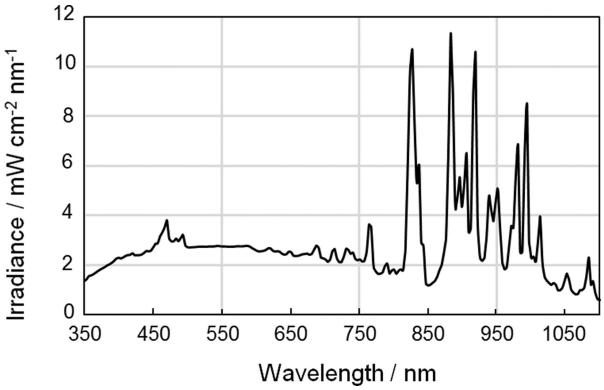
Extended Data Fig. 3 | Time course of H_2 and O_2 evolution over $SrTiO_3$:Al photodeposited with Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%) under simulated sunlight irradiation.



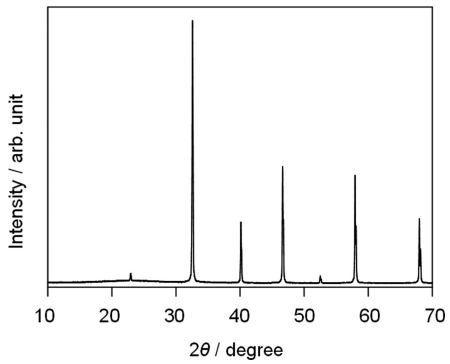
Extended Data Fig. 4 | Effect of initial background pressure on the photocatalytic water-splitting activity of $SrTiO_3$:Al photodeposited with Rh (0.1 wt%)/Cr_2O_3 (0.05 wt%)/CoOOH (0.05 wt%).



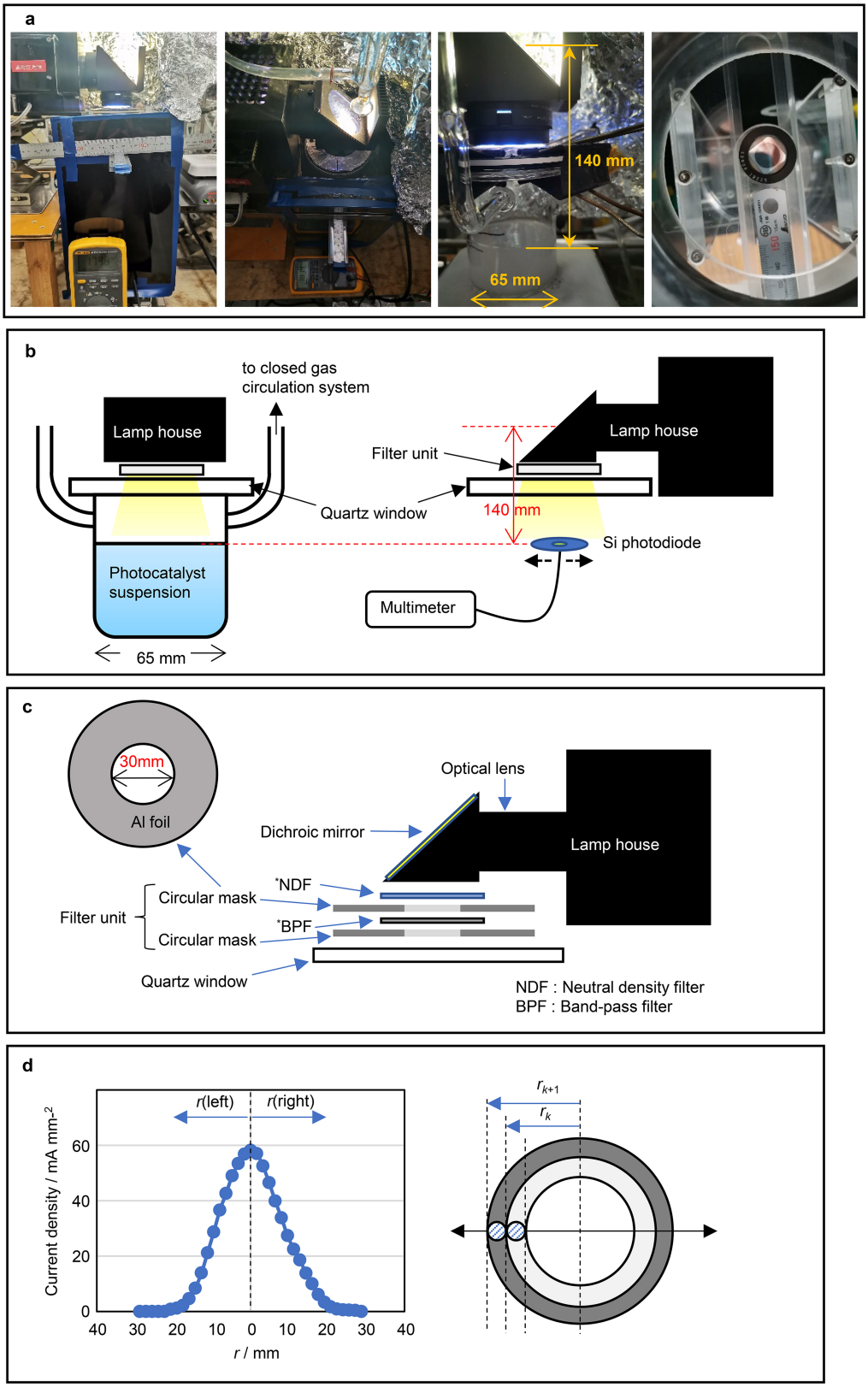
Extended Data Fig. 5 | Simulations of graded band energy values and resulting anisotropic charge separation based on various differences in work function between the {110} and {100} facets. a, Maps of conduction band energy and logarithms of the electron and hole densities (*n* and *p*, respectively). **b**, **c**, Energy band diagrams (**b**) and electron and hole densities (**c**) as functions of position. See Methods for further discussion.



Extended Data Fig. 6 | Power spectrum of the Xe lamp used in the study.



Extended Data Fig. 7 | XRD pattern of the synthesized $SrTiO_3$:Al.



Extended Data Fig. 8 | EQE measurement. a, Photographs of the devices used in the measurement. Left, side view of the measurement system. Middle left, top view of the measurement system. Middle right, arrangement of the lamp and reactor. Right, arrangement of the Si photodiode. **b**, **c**, Illustrations of water-splitting reactor and photon-counting system (**b**) and illumination unit (**c**). **d**, Example of light intensity distribution (365-nm bandpass filter and 0.1-optical-density neutral density filter; left) and model used to calculate the number of incident photons (right).

Extended Data Table 1 | Results of EQE measurements

Wavelength / nm	OD#	Photons / h	Activity / μmol h^{-1}		EQE/%
			H_2	O_2	
350	0	5.26×10^{19}	41.8	20.3	95.7
360	0	1.27×10^{20}	101	54.3	95.9
365	0	1.36×10^{21}	1060	512	93.2
365	0.3	7.49×10^{20}	557	276	89.4
365	0.5	4.84×10^{20}	366	177	91.1
365	1.0	1.70×10^{20}	131	67.3	92.8
370	0	2.54×10^{20}	126	65.2	59.7
380	0	2.06×10^{20}	57.3	25.8	33.6

The EQE value at 365 nm was essentially independent of the light intensity over the range of intensities examined. The average EQE in this series of experiments was 91.6%, which was used to generate the plot shown in Fig. 1b.

#OD, optical density.

Extended Data Table 2 | Parameters used in the calculations of the energy band diagrams and electron and hole densities for the $SrTiO_3$:Al particulate system

Parameters	Numerical values
Effective density of states - conduction band	7.94×10^{20} cm^{-3} (ref.25) [a]
Effective density of states - valence band	6.71×10^{20} cm^{-3} [a]
Charge carrier mobility	6 $cm^2V^{-1}s^{-1}$ (ref.25)
Dielectric constant	300 (ref.26)
Depth of Fermi level from Conduction band	0.5 eV (ref.27)
Band Gap	3.2 eV (ref.27) [b]
Electron affinity	4.2 eV (ref. 28)
Absorption Coefficient (α) at 365 nm	1.3×10^4 cm^{-1} (ref.29,30)
The number of incident photon (N) at 365 nm	9.55×10^{15} $cm^{-2}s^{-1}$ [b]
Generation rate ($G = \alpha N$)	1.24×10^{20} $cm^{-3}s^{-1}$
Charge carrier lifetime	24 ns [c]

[a]The effective densities of states for the conduction and valence bands were obtained from the reported effective mass.

[b]The numerical values of these parameters were obtained from the experimental observations and conditions.

[c]The charge carrier lifetime used for trap-assisted Shockley–Reed–Hall recombination is calculated so that the diffusion length of the photogenerated charge carriers is greater than the particle dimensions.

Data from refs. [25–30].

Article

A general carbonyl alkylative amination for tertiary amine synthesis

https://doi.org/10.1038/s41586-020-2213-0

Received: 30 October 2019

Accepted: 27 March 2020

Published online: 8 April 2020

Check for updates

Roopender Kumar[1], Nils J. Flodén[1,2], William G. Whitehurst[1,2] & Matthew J. Gaunt[1]✉

The ubiquity of tertiary alkylamines in pharmaceutical and agrochemical agents, natural products and small-molecule biological probes[1,2] has stimulated efforts towards their streamlined synthesis[3–9]. Arguably the most robust method for the synthesis of tertiary alkylamines is carbonyl reductive amination[3], which comprises two elementary steps: the condensation of a secondary alkylamine with an aliphatic aldehyde to form an all-alkyl-iminium ion, which is subsequently reduced by a hydride reagent. Direct strategies have been sought for a 'higher order' variant of this reaction via the coupling of an alkyl fragment with an alkyl-iminium ion that is generated in situ[10–14]. However, despite extensive efforts, the successful realization of a 'carbonyl alkylative amination' has not yet been achieved. Here we present a practical and general synthesis of tertiary alkylamines through the addition of alkyl radicals to all-alkyl-iminium ions. The process is facilitated by visible light and a silane reducing agent, which trigger a distinct radical initiation step to establish a chain process. This operationally straightforward, metal-free and modular transformation forms tertiary amines, without structural constraint, via the coupling of aldehydes and secondary amines with alkyl halides. The structural and functional diversity of these readily available precursors provides a versatile and flexible strategy for the streamlined synthesis of complex tertiary amines.

Carbonyl reductive amination is an effective method for the preparation of linear tertiary alkylamines; however, the synthesis of branched variants is more problematic[3]. The condensation of a secondary alkylamine with a di-alkylketone is often slow and may require the use of activating reagents; di-alkylketones are also not as readily available as aldehydes and may require multi-step syntheses. A method by which an alkyl group could be directly added to an aldehyde-derived alkyl-iminium ion would circumvent these problems and would increase complexity by leveraging three, rather than two, programmable feedstocks into a tertiary alkylamine synthesis. Unfortunately, the most logical approach to this multi-component strategy—the direct addition of common organometallic nucleophiles to alkyl-iminium ions—is seldom successful in delivering the tertiary alkylamine product[11–14]. Organometallic reagents—such as Grignard and alkyl-lithium reagents—are rarely compatible with in-situ iminium ion formation from alkyl aldehydes and secondary alkylamines, necessitating the pre-formation and isolation of an unstable iminium ion[12,15]. Additionally, the high basicity of these organometallic reagents results in competitive deprotonation of the C–H bond adjacent to the carbon–nitrogen double bond of the iminium ion, substantially restricting the scope of the process[16]. Less reactive zinc- or cerium-based organometallic compounds also exhibit limited scope and are restricted to activated alkyl fragments[12,13,15,17,18]. The Petasis reaction offers broader scope in the amine component, but has specific substrate requirements for the carbonyl and organoboron-derived components[19]. More efficient reactivity has been demonstrated with C(sp)-nucleophiles[20,21], C(sp^2)-nucleophiles[19,22] and allyl-nucleophiles[19,23], or with activating auxiliary-derived imines[4]; however, the successful deployment of a suitably reactive and generally available source of unactivated alkyl nucleophile for addition to an alkyl-iminium ion continues to prove challenging. Considering the limitations of the organometallic methods, strategies based on the addition of neutral alkyl-radical species to imine-derivatives have also been investigated[10,24–26]. However, the low electrophilicity of the carbon–nitrogen double bond means that the use of an activating group either on the nitrogen atom (R^1) or in the carbonyl component (R^2) is essential to render the imine-derivative sufficiently reactive towards alkyl-radicals, thereby limiting the scope and practical application of these reactions with respect to the downstream synthesis of tertiary amine targets[10,24–26] (Fig. 1b).

In our efforts to develop a direct carbonyl alkylative amination (CAA) method, we proposed that the addition of a neutral carbon-centred radical to an in-situ-generated, positively charged, alkyl-iminium ion would obviate the requirement for auxiliary-activated imines and provide a single-step synthesis of tertiary alkylamines. To the best of our knowledge, the elementary step comprising the direct intermolecular addition of an alkyl-radical to an alkyl-iminium ion has not been reported, with the exception of one example from 1991, in which an organo-mercury salt was irradiated with ultraviolet (UV) light to generate an alkyl-radical that was added to a formaldehyde-derived iminium ion[27]. Addition of a neutral nucleophilic alkyl-radical to an alkyl-iminium ion would afford an aminium radical cation, a species that could then be intercepted through hydrogen atom transfer to form the protonated tertiary alkylamine (Fig. 1c). A radical-based CAA requires the orchestration of a number of simultaneously occurring events to

[1]Department of Chemistry, University of Cambridge, Cambridge, UK. [2]These authors contributed equally: Nils J. Flodén, William G. Whitehurst. ✉e-mail: mjg32@cam.ac.uk

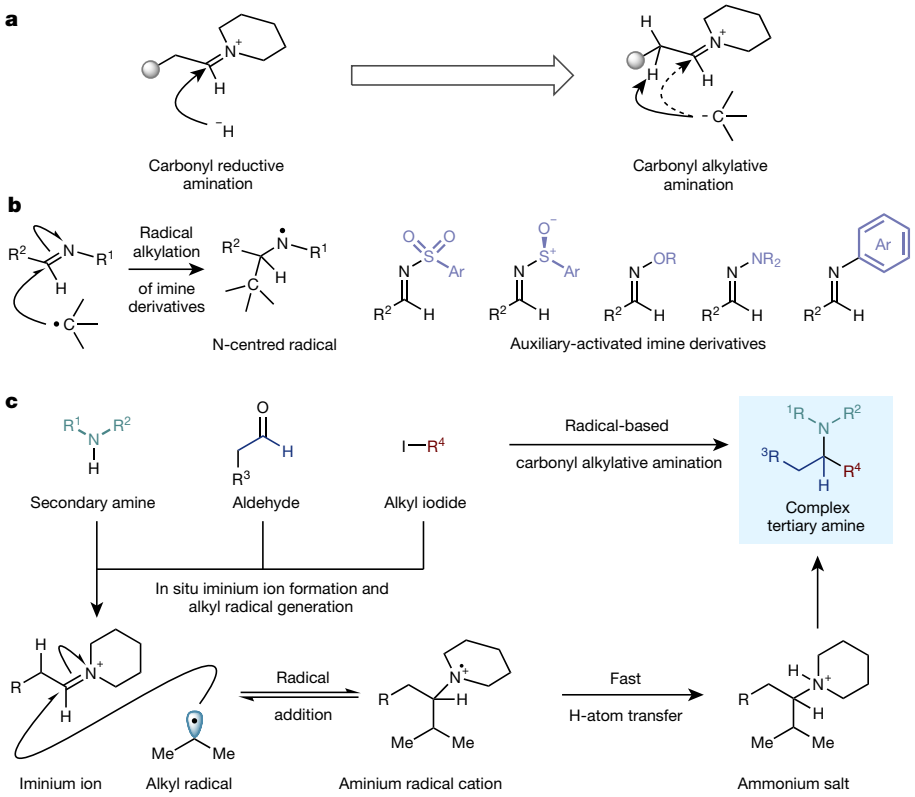
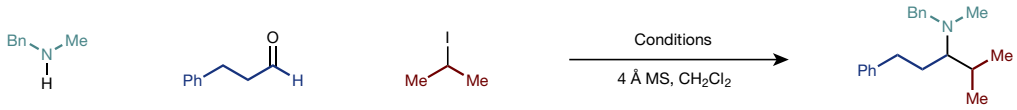


Fig. 1 | Evolution of a strategy for carbonyl alkylative amination. a, Addition of alkyl groups to alkyl-iminium ions remains a challenging transformation. **b**, Radical alkylation of auxiliary-activated imine derivatives. **c**, Carbonyl alkylative amination for the synthesis of complex tertiary alkylamines: the addition of alkyl radicals to alkyl-iminium ions.

form reactive intermediates, each of which is capable of following competitive and deleterious reaction pathways. For example, it is necessary to maintain a high concentration of the reactive alkyl-iminium ion in order to effectively engage the incipient alkyl-radical. Furthermore, the reagent effecting the hydrogen atom transfer must be capable of rapidly intercepting the transient aminium radical cation but must not reduce the iminium ion. Here we report the realization of our hypothesis through the development of modular and efficient CAA, a method that combines three abundant feedstocks—secondary amine, aldehyde and alkyl-halide—in a single step. A key part of this strategy is the use of visible light to facilitate a radical initiation step under mild conditions, leading to a practical and general synthesis of complex tertiary alkylamines.

Our initial investigations focused on a representative reaction between *N*-methylbenzylamine (**1a**), hydrocinnamaldehyde (**2a**) and 2-iodopropane (**3a**) (Fig. 2a). Attempts to achieve CAA using classical protocols for radical generation failed to produce an efficient reaction[28]. For example, the use of tributyltin hydride (Bu_3Sn–H) and azobisisobutyronitrile (AIBN) produced trace amounts of alkylamine **4a** (entry 1); this could be modestly improved by using tris(trimethylsilyl) silane ($(Me_3Si)_3Si$–H) in place of the tin reagent (entry 2)[29]. We evaluated the use of different additives in order to promote the high concentrations of alkyl-iminium required to effectively intercept the alkyl-radical. High conversion was achieved in the presence of *tert*-butyldimethylsilyl trifluoromethanesulfonate (TBSOTf), as determined by ^{1}H nuclear magnetic resonance (NMR) spectroscopy (Supplementary Information 1). For example, a thermal reaction combining $(Me_3Si)_3Si$–H, AIBN and TBSOTf produced the desired alkylamine **4a**, but as a 3:1 mixture with the corresponding reductive amination product **4a′** (entry 3; **4a′** not shown). Notably, a reaction that combined TBSOTf with Bu_3Sn–H and AIBN under thermal conditions resulted exclusively in reductive amination to **4a′** (entry 4). Further exploration of the target transformation (Supplementary Information 1) led us to identify a set of unique and simpler reaction conditions that exploited the effect of visible-light activation; after parameter optimization, this approach produced high yields of the desired alkylamine **4a** (92% assay yield and 80% isolated yield) with only trace amounts of reductive amination (entry 5). When TBSOTf was omitted, **4a** was still obtained in 82% assay yield, providing a set of conditions that could be used when acid-sensitive functional groups were present. An important aspect of these mild reaction conditions is the distinct nature of the radical initiation step. Although there are a number of possible explanations for this phenomenon, homolysis of the C–I bond seems unlikely given that a control reaction using a 455-nm long-pass filter—which removes the minor UV and near-UV components of the blue LED lamp that are required for the homolysis[30,31]—still generated **4a** in 86% yield (entry 7). We then systematically studied the light-absorbing properties of each component and of likely intermediates involved in the CAA reaction—secondary amine, alkyl-aldehyde, alkyl-iodide, enamine, iminium and $(Me_3Si)_3Si$–H—as well as their combinations, in order to investigate the possibility of photosensitization or an electron donor–acceptor complex that could be involved in radical generation[32]. Although neither any component alone nor the combination of any two components showed absorption above or close to 455 nm, the UV–visible absorption spectrum of a ternary mixture comprising enamine (formed by condensation of **1a** and **2a**), 2-iodopropane (**3a**) and $(Me_3Si)_3Si$–H revealed a new, red-shifted band (400–500 nm). Although at present we do not have sufficient evidence to invoke a specific interaction, this observation potentially supports a multi-component interaction, which leads to radical initiation upon excitation with visible light (Supplementary Information 1). In any case, this radical initiation process offers distinct advantages over other light-mediated processes: whereas previous reports have been restricted to the formation of carbon–heteroatom bonds involving the initiating reagent[31,33]; in this process the $(Me_3Si)_3Si$–H is not

a

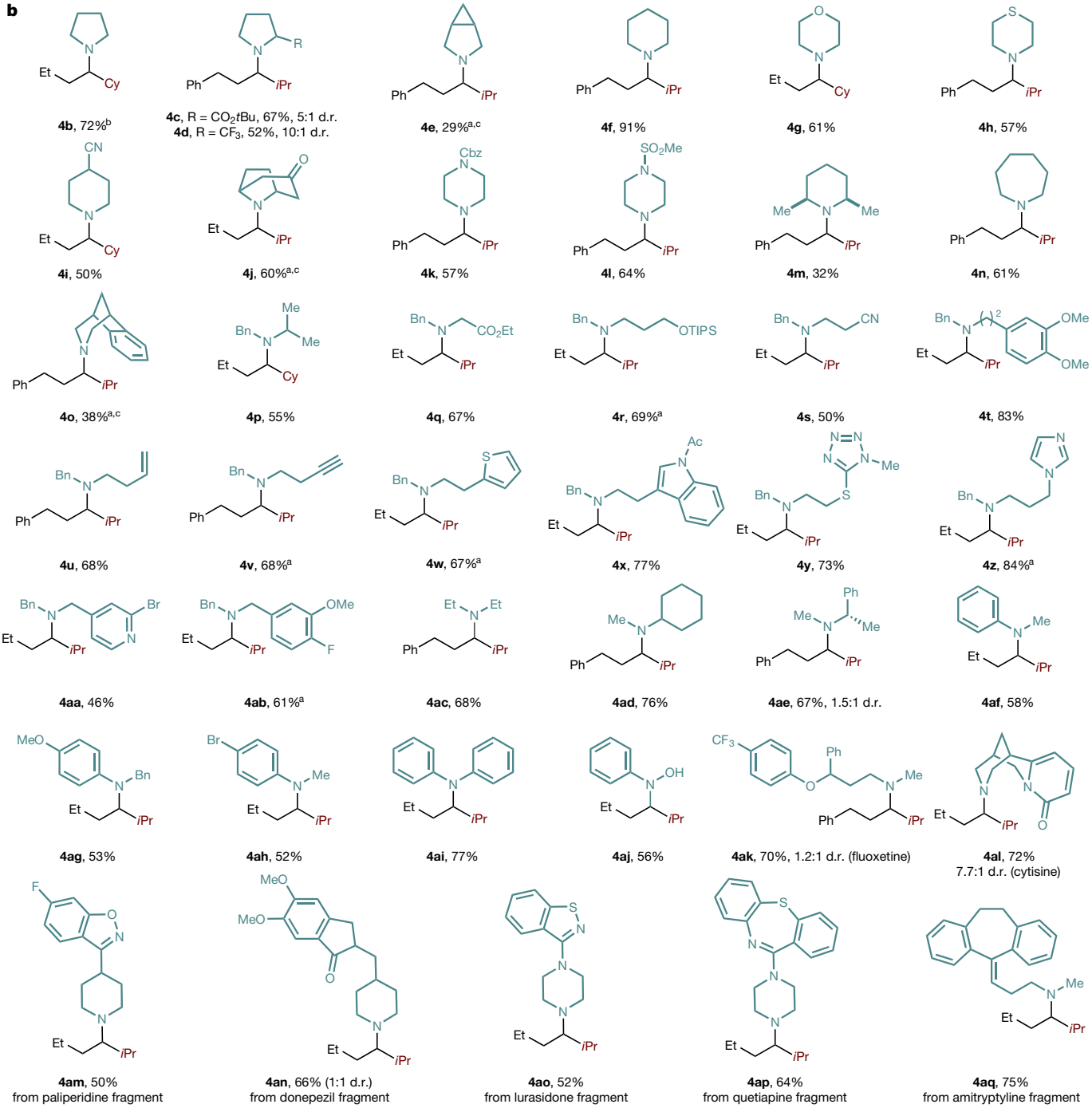


Secondary amine **1a** | Aldehyde **2a** (2 equiv.) | Alkyl iodide **3a** (3 equiv.) | **4a**

Conditions	Yield of **4a** by ^{1}H NMR
1. Bu_3Sn–H, AIBN, 70 °C	Trace
2. $(Me_3Si)_3Si$–H, AIBN, 70 °C	<10%
3. $(Me_3Si)_3Si$–H, AIBN, TBSOTf, 70 °C	75% (plus 25% reductive amination)
4. Bu_3Sn–H, AIBN, TBSOTf, 70 °C	0% (84% reductive amination)

Conditions	Yield of **4a** by ^{1}H NMR
5. 2 equiv. $(Me_3Si)_3Si$–H, 1 equiv. TBSOTf, RT, 8 h 40 W blue LED	92% (80% isolated)
6. $(Me_3Si)_3Si$–H, RT, 40 W blue LED	82%
7. $(Me_3Si)_3Si$–H, TBSOTf, RT, 40 W blue LED, 455 nm filter	86%

b



4b, 72%[b]

4c, R = CO_2tBu, 67%, 5:1 d.r.
4d, R = CF_3, 52%, 10:1 d.r.

4e, 29%[a,c]

4f, 91%

4g, 61%

4h, 57%

4i, 50%

4j, 60%[a,c]

4k, 57%

4l, 64%

4m, 32%

4n, 61%

4o, 38%[a,c]

4p, 55%

4q, 67%

4r, 69%[a]

4s, 50%

4t, 83%

4u, 68%

4v, 68%[a]

4w, 67%[a]

4x, 77%

4y, 73%

4z, 84%[a]

4aa, 46%

4ab, 61%[a]

4ac, 68%

4ad, 76%

4ae, 67%, 1.5:1 d.r.

4af, 58%

4ag, 53%

4ah, 52%

4ai, 77%

4aj, 56%

4ak, 70%, 1.2:1 d.r. (fluoxetine)

4al, 72%
7.7:1 d.r. (cytisine)

4am, 50%
from paliperidine fragment

4an, 66% (1:1 d.r.)
from donepezil fragment

4ao, 52%
from lurasidone fragment

4ap, 64%
from quetiapine fragment

4aq, 75%
from amitryptyline fragment

Fig. 2 | Scope of the amine component in carbonyl alkylative amination. **a**, Optimal reaction conditions. MS, molecular sieves. **b**, Scope of the amine component. [a]No TBSOTf used. [b]TMSOTf (trimethylsilyl trifluoromethanesulfonate) used instead of TBSOTf. [c]Amine·HCl salt used. Cbz, benzyloxycarbonyl; Cy, cyclohexyl; TIPS, triisopropylsilyl; d.r., diastereomeric ratio.

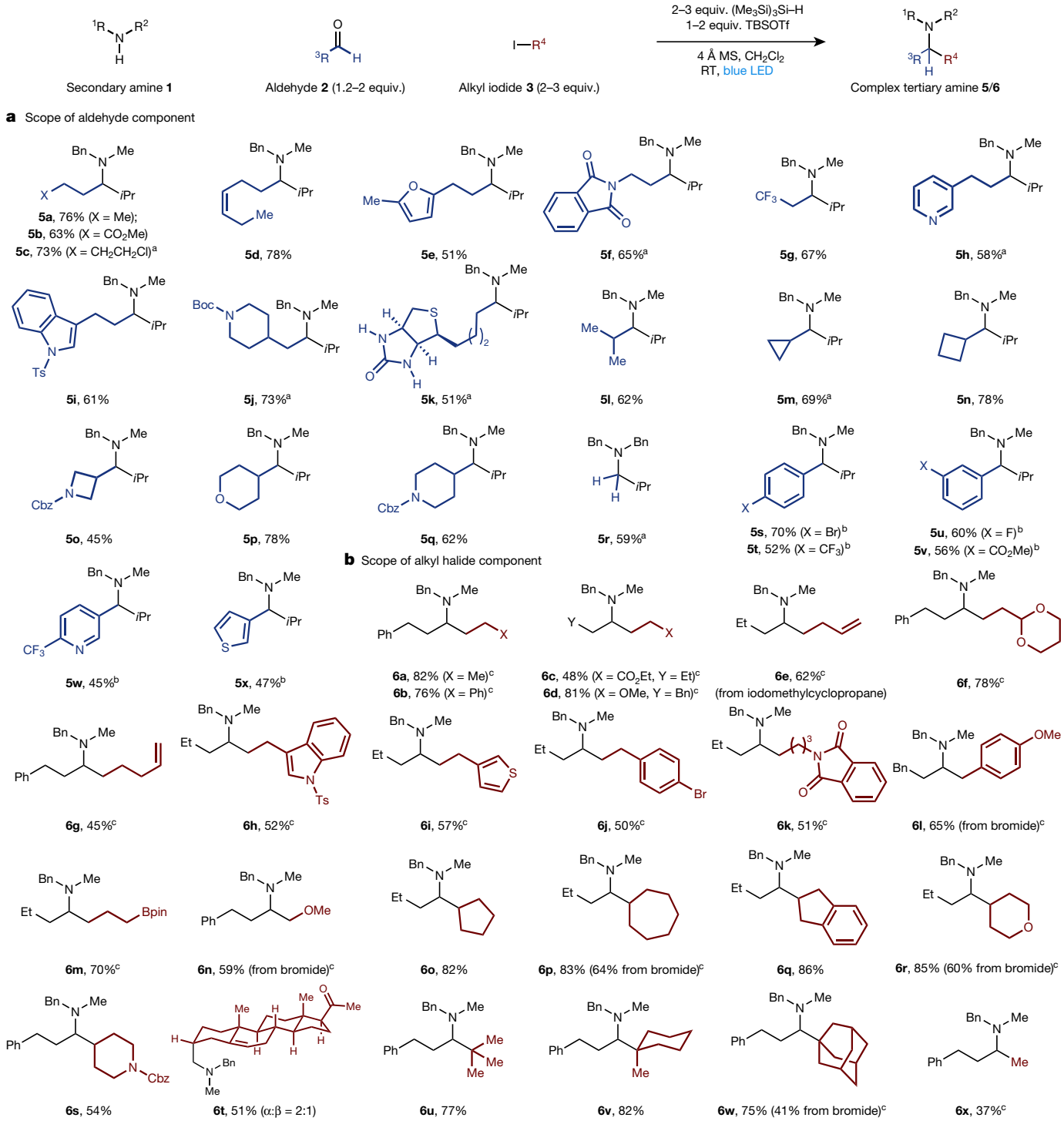


Fig. 3 | Scope of the carbonyl alkylative amination reaction. a, Scope of the aldehyde component. **b**, Scope of the alkyl halide component. [a]No TBSOTf used. [b]TMSOTf used instead of TBSOTf. [c]5–10 mol% 2-methyl-2-iodopropionate used as an initiator.

incorporated in the newly formed bond, thus enabling the generation of a C(sp^3)–C(sp^3) bond. Furthermore, the benign nature of the new radical generation step, in the absence of classical chemical initiators, could aid the development of straightforward, metal-free radical C(sp^3)–C(sp^3) bond-forming reactions from alkyl-halides.

Having established a viable protocol, we began an extensive investigation of the scope of CAA by first testing its capacity to produce α-branched cyclic tertiary alkylamines. A wide range of functionalized saturated cyclic and heterocyclic secondary alkylamines could be successfully used in this radical-based process (Fig. 2b), producing the desired alkylamines (**4b**–**4o**). Notably, auxiliary-activated radical-addition methods cannot readily access this class of tertiary alkylamines, which can be produced in a single step using the new protocol. We found that *N*,*N*-dialkylamines with a range of both linear and branched functionalized alkyl substituents—including aromatic heterocycles—gave the tertiary alkylamines **4a**, **4p**–**4ae** in good yields.

For certain amines with electron-withdrawing groups close to the nitrogen atom, reductive amination was observed as a competing side reaction (10% with **4s**). *N*-alkyl anilines and even poorly nucleophilic diarylamines could be used effectively (**4af**–**4ai**), which expands the

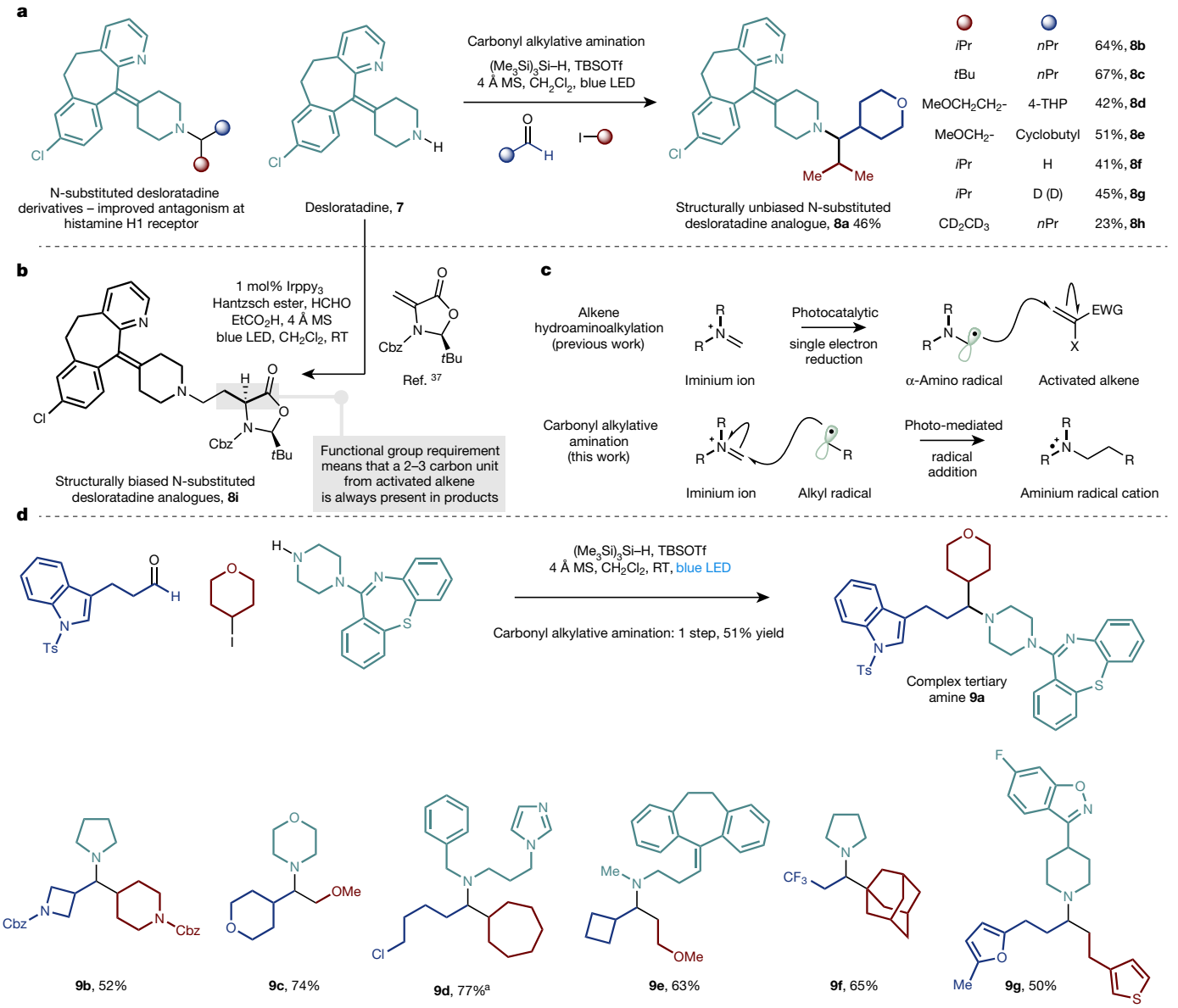


Fig. 4 | One-step synthesis of complex tertiary alkylamines via carbonyl alkylative amination and its comparison with related methods. a, Rapid and streamlined synthesis of desloratadine analogues by CAA. **b**, An example of amine synthesis using photocatalytic alkene hydroaminoalkylation, highlighting the 'biased' nature of the products resulting from the activated alkene acceptor. EWG, electron withdrawing group. **c**, The distinct mechanistic pathways of alkene hydroaminoalkylation and CAA. **d**, Structurally diverse tertiary alkylamines formed by CAA in one step from readily available building blocks. [a]No TBSOTf used.

potential scope of the CAA process as these functionalities are ubiquitous in pharmaceutically relevant molecules. A reaction using *N*-phenyl hydroxylamine produced the expected product (**4aj**) via addition to the corresponding nitrone intermediate. To demonstrate the compatibility of this process with features that are typically encountered in pharmaceutical agents, we showed that a range of drug fragments could tolerate the reaction conditions, resulting in the desired complex tertiary amine products in synthetically useful yields (**4ak**–**4aq**).

Using a selection of functionalized linear aldehydes, we were able to prepare the corresponding tertiary alkylamines (Fig. 3a, **5a**–**5k**). Branched aldehydes—including those with saturated cyclic and heterocyclic rings—also performed well, forming the hindered tertiary alkylamines **5l**–**5q**. Beyond aliphatic aldehydes, we found that formaldehyde (to **5r**), substituted benzaldehydes (to **5s**–**5v**) and heteroaryl aldehydes (**5w**–**5x**) function well in the reaction, and their successful conversion to amine products considerably expands the scope of CAA.

In the case of primary alkyl iodides, the reaction was aided by the addition of 5–10 mol% of the radical chain initiator ethyl 2-iodo-2-methylpropionate[34], which can generate alkyl radicals more easily than can primary alkyl iodides. Both simple and functionalized primary alkyl iodides proved to be good coupling partners under these conditions, with linear alkane fragments added to alkyl-iminium ions to form tertiary alkylamines **6a**–**6n**. Although the reaction could accommodate various functional groups in the alkyl-halide, slightly lower yields were observed for radicals containing proximal electron withdrawing groups (**6c**). This is likely to originate from a competitive hydrogen atom transfer between the silane and the alkyl-radical, as well as the slower rate of addition of a less nucleophilic alkyl-radical to the iminium ion. Benzyl bromides and methoxymethyl bromide, the iodides of which are unstable, were found to be suitable alkylating agents and gave the amine products in synthetically useful yields (**6l**, **6n**). As expected, secondary alkyl-halides were competent substrates

and formed amines **6o**–**6t** in good yields. Tertiary alkyl-iodides were also excellent coupling partners: the hindered alkylamine products **6u**–**6w**, which would be challenging to prepare via reductive amination, were obtained in high yield. A selection of unactivated alkyl-bromides reacted under the modified conditions—with the radical chain initiator[34]—to give reasonable yields of the tertiary alkylamines **6p**, **6r** and **6w**; this provides an alternative protocol if an alkyl-iodide is not available. We were pleased to find that iodomethane, acting as a source of methyl-radical, could be added to an iminium ion to provide access to an important class of branched tertiary alkylamines (**6x**). In addition to the 90 examples of successful CAA documented in Figs. 2 and 3, we have detailed an extended assessment of the reaction scope (Supplementary Fig. 9, Supplementary Information 1) that includes substrates that give moderate—but still synthetically useable—yields, as well as examples for which the process is low yielding or unsuccessful.

As a further demonstration of this methodology, we showed that the pharmaceutical agent desloratadine, which contains a secondary amine, can be used as the amine component of this CAA to directly append two interchangeable molecular fragments within the α-branched tertiary alkylamine product (Fig. 4a). Tertiary amine derivatives of desloratadine have recently attracted attention owing to their improved pharmacokinetic properties, and convey a longer duration of antagonism at the histamine H1 receptor[35]. Using the standard coupling conditions, we showed that a collection of desloratadine-derived α-branched tertiary amine products can be readily prepared in a modular fashion (**8a**–**g**). As a demonstration of the efficacy of these reactions, we prepared an α-d_2-*N*-alkyl derivative of desloratadine (>95% deuterium incorporation, **8g**) in a single step from commercial d_2-paraformaldehyde and 2-iodopropane. Furthermore, a similar reaction using d_5-iodoethane and a linear aldehyde provided direct access to a different class of labelled tertiary amine (**8h**)[36].

Direct CAA on desloratadine is distinct from our previous work[37] on photoredox-mediated alkylamine synthesis, which was based on alkene hydroaminoalkylation via the generation and reaction of α-amino radicals (Fig. 4b, c). Although we previously showed that this secondary amine can engage activated alkenes (to **8i**), the tertiary alkylamine products prepared by that approach always contain the structural signature derived from the intrinsic requirements of a reaction that necessitates use of an activated alkene acceptor. By contrast, the present mechanistically distinct reaction enables the use of a wide range of primary, secondary and tertiary alkyl-halides, providing direct access to tertiary alkylamines that are unbiased by any functional requirements of the transformation and contain no trace of the activating groups required in the readily available building blocks.

A major advantage of this CAA strategy is its modularity; the vast array of coupling partners within each of the three abundant feedstocks required for this reaction means that structural and functional diversity can be easily programmed into the tertiary amine products (Fig. 4d). To illustrate this, we evaluated the cross-compatibility of the process by varying each reaction component in order to produce a range of complex tertiary alkylamines with diverse structural and functional properties. The robust nature of the reaction is reflected by the ease with which densely functionalized alkylamines **9a**–**9g** can be produced in a single step from readily available building blocks, highlighting the suitability of such alkylamines for early-stage drug discovery applications (Fig. 4d). The streamlined nature by which these products are prepared demonstrates the synthetic potential of this process.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2213-0.

1. Roughley, S. D. & Jordan, A. M. The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **54**, 3451–3479 (2011).
2. Blakemore, D. C. et al. Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **10**, 383–394 (2018).
3. Abdel-Magid, A. F. & Mehrman, S. J. A review on the use of sodium triacetoxyborohydride in the reductive amination of ketones and aldehydes. *Org. Process Res. Dev.* **10**, 971–1031 (2006).
4. Robak, M. T., Herbage, M. A. & Ellman, J. A. Synthesis and applications of *tert*-butanesulfinamide. *Chem. Rev.* **110**, 3600–3740 (2010).
5. Huang, L., Arndt, M., Gooßen, K., Heydt, H. & Gooßen, L. J. Late transition metal-catalyzed hydroamination and hydroamidation. *Chem. Rev.* **115**, 2596–2697 (2015).
6. Pirnot, M. T., Wang, Y.-M. & Buchwald, S. L. Copper hydride catalyzed hydroamination of alkenes and alkynes. *Angew. Chem. Int. Ed.* **55**, 48–57 (2016).
7. Musacchio, A. J. et al. Catalytic intermolecular hydroaminations of unactivated olefins with secondary alkyl amines. *Science* **355**, 727–730 (2017).
8. Matier, C. D., Schwaben, J., Peters, J. C. & Fu, G. C. Copper-catalyzed alkylation of aliphatic amines induced by visible light. *J. Am. Chem. Soc.* **139**, 17707–17710 (2017).
9. Grogan, G. Synthesis of chiral amines using redox biocatalysis. *Curr. Opin. Chem. Biol.* **43**, 15–22 (2018).
10. Friestad, G. K. Addition of carbon-centered radicals to imines and related compounds. *Tetrahedron* **57**, 5461–5496 (2001).
11. Reiber, H. G. & Stewart, T. D. The tetra alkyl methylene immonium salts. *J. Am. Chem. Soc.* **62**, 3026–3030 (1940).
12. Paukstelis, J. W. & Cook, A. G. in *Enamines: Synthesis, Structure and Reactions* 2nd edn (ed. Cook, A. G.) Ch. 6, 275–346 (Marcel Dekker, 1988).
13. Bloch, R. Additions of organometallic reagents to C=N bonds: reactivity and selectivity. *Chem. Rev.* **98**, 1407–1438 (1998).
14. Kaiser, E. M. Lithium: annual survey covering the year 1975. *J. Organomet. Chem.* **130**, 1–131 (1977).
15. Werner, V., Ellwart, M., Wagner, A. J. & Knochel, P. Preparation of tertiary amines by the reaction of iminium ions derived from unsymmetrical aminals with zinc and magnesium organometallics. *Org. Lett.* **17**, 2026–2029 (2015).
16. Saidi, M. R. & Nazari, M. Aminoalkylation with aldehydes mediated by solid lithium perchlorate. *Monatsh. Chem.* **135**, 309–312 (2004).
17. Agosti, A., Britto, S. & Renaud, P. An efficient method to convert lactams and amides into 2,2-dialkylated amines. *Org. Lett.* **10**, 1417–1420 (2008).
18. Haurena, C., LeGall, E., Sengmany, S. & Martens, T. Chiral amines in the diastereoselective Mannich-related multicomponent synthesis of diarylmethylamines, 1,2-diarylethylamines, and β-arylethylamines. *Tetrahedron* **66**, 9902–9911 (2010).
19. Wu, P., Givskov, M. & Nielsen, T. E. Reactivity and synthetic applications of multicomponent Petasis reactions. *Chem. Rev.* **119**, 11245–11290 (2019).
20. Aschwanden, P. & Carreira, E. M. in *Acetylene Chemistry: Chemistry, Biology and Material Science* (eds Diederich, F., Stang, P. J. & Tywinski, R. R.) (Wiley, 2005).
21. Lauder, K., Toscani, A., Scalacci, N. & Castagnolo, D. Synthesis and reactivity of propargylamines in organic chemistry. *Chem. Rev.* **117**, 14091–14200 (2017).
22. Heinz, C. et al. Ni-catalyzed carbon–carbon bond-forming reductive amination. *J. Am. Chem. Soc.* **140**, 2292–2300 (2018).
23. Chen, T.-Y., Tsutsumi, R., Montgomery, T. P., Volchkov, I. & Krische, M. J. Ruthenium-catalyzed C–C coupling of amino alcohols with dienes via transfer hydrogenation: redox-triggered imine addition and related hydroaminoalkylations. *J. Am. Chem. Soc.* **137**, 1798–1801 (2015).
24. Miyabe, H., Yoshioka, E. & Kohtani, S. Progress in intermolecular carbon radical addition to imine derivatives. *Curr. Org. Chem.* **14**, 1254–1264 (2010).
25. Friestad, G. K. Radical additions to chiral hydrazones: stereoselectivity and functional group compatibility. *Top. Curr. Chem.* **320**, 61–91 (2011).
26. Tauber, J., Imbri, D. & Opatz, T. Radical addition to iminium ions and cationic heterocycles. *Molecules* **19**, 16190–16222 (2014).
27. Russell, G. A., Yao, C.-F., Rajaratnam, R. & Kim, B. H. Promotion of electron transfer by protonation of nitrogen-centered free radicals. The addition of radicals to iminium ions. *J. Am. Chem. Soc.* **113**, 373–375 (1991).
28. Baguley, P. A. & Walton, J. C. Flight from the tyranny of tin: the quest for practical radical sources free from metal encumbrances. *Angew. Chem. Int. Ed.* **37**, 3072–3082 (1998).
29. Chatgilialoglu, C., Ferreri, C., Landais, Y. & Timokhin, V. I. Thirty years of $(TMS)_3SiH$: a milestone in radical-based synthetic chemistry. *Chem. Rev.* **118**, 6516–6572 (2018).
30. Le, C., Chen, T. Q., Liang, T., Zhang, P. & MacMillan, D. W. C. A radical approach to the copper oxidative addition problem: trifluoromethylation of bromoarenes. *Science* **360**, 1010–1014 (2018).
31. Cheng, Y., Mück-Lichtenfeld, C. & Studer, A. Metal-free radical borylation of alkyl and aryl iodides. *Angew. Chem. Int. Ed.* **57**, 16832–16836 (2018).
32. Bahamonde, A. & Melchiorre, P. Mechanism of the stereoselective α-alkylation of aldehydes driven by the photochemical activity of enamines. *J. Am. Chem. Soc.* **138**, 8019–8030 (2016).
33. Fawcett, A. et al. Photoinduced decarboxylative borylation of carboxylic acids. *Science* **357**, 283–286 (2017).
34. Lei, L. et al. Systematic study on alkyl iodide initiators in living radical polymerization with organic catalysts. *Macromolecules* **47**, 6610–6618 (2014).
35. Bosma, R. et al. Route to prolonged residence time at the histamine H_1 receptor: growing from desloratadine to rupatadine. *J. Med. Chem.* **62**, 6630–6644 (2019).
36. Atzrodt, J., Derdau, V., Kerr, W. J. & Reid, M. Deuterium- and tritium-labelled compounds: applications in the life sciences. *Angew. Chem. Int. Ed.* **57**, 1758–1784 (2018).
37. Trowbridge, A., Reich, D. & Gaunt, M. J. Multicomponent synthesis of tertiary alkylamines by photocatalytic olefin-hydroaminoalkylation. *Nature* **561**, 522–527 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Data availability

Materials and methods, experimental procedures, useful information, mechanistic studies, optimization studies, ^{1}H NMR spectra, ^{13}C NMR spectra and mass spectrometry data are available in the Supplementary Information. Raw data are available from the corresponding author on reasonable request.

Acknowledgements We acknowledge the Swiss National Science Foundation (R.K.), the Gates Cambridge Trust (N.J.F.), the EPSRC (W.G.W.) and the Royal Society (for a Wolfson Merit Award, M.J.G.).

Author contributions R.K., N.J.F., W.G.W. and M.J.G. conceived the project; R.K., N.J.F. and W.G.W. conducted and analysed the experiments; and R.K., N.J.F., W.G.W. and M.J.G. wrote the manuscript. N.J.F. and W.G.W. contributed equally to the project and are listed alphabetically.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2213-0.

Correspondence and requests for materials should be addressed to M.J.G.

Reprints and permissions information is available at http://www.nature.com/reprints.

Article

Skeleton of a Cretaceous mammal from Madagascar reflects long-term insularity

https://doi.org/10.1038/s41586-020-2234-8

Received: 5 November 2019

Accepted: 26 February 2020

Published online: 29 April 2020

Check for updates

David W. Krause[1,2]✉, Simone Hoffmann[3], Yaoming Hu[2,15], John R. Wible[4], Guillermo W. Rougier[5], E. Christopher Kirk[6,7], Joseph R. Groenke[2,8], Raymond R. Rogers[9], James B. Rossie[10], Julia A. Schultz[11], Alistair R. Evans[12,13], Wighart von Koenigswald[11] & Lydia J. Rahantarisoa[14]

The fossil record of mammaliaforms (mammals and their closest relatives) of the Mesozoic era from the southern supercontinent Gondwana is far less extensive than that from its northern counterpart, Laurasia[1,2]. Among Mesozoic mammaliaforms, Gondwanatheria is one of the most poorly known clades, previously represented by only a single cranium and isolated jaws and teeth[1–5]. As a result, the anatomy, palaeobiology and phylogenetic relationships of gondwanatherians remain unclear. Here we report the discovery of an articulated and very well-preserved skeleton of a gondwanatherian of the latest age (72.1–66 million years ago) of the Cretaceous period from Madagascar that we assign to a new genus and species, *Adalatherium hui*. To our knowledge, the specimen is the most complete skeleton of a Gondwanan Mesozoic mammaliaform that has been found, and includes the only postcranial material and ascending ramus of the dentary known for any gondwanatherian. A phylogenetic analysis including the new taxon recovers Gondwanatheria as the sister group to Multituberculata. The skeleton, which represents one of the largest of the Gondwanan Mesozoic mammaliaforms, is particularly notable for exhibiting many unique features in combination with features that are convergent on those of therian mammals. This uniqueness is consistent with a lineage history for *A. hui* of isolation on Madagascar for more than 20 million years.

Island environments promote evolutionary trajectories among mammals and other vertebrates that contrast with those on continents, and which result in demonstrable anatomical, physiological and behavioural differences[6–10]. These differences have previously been ascribed to markedly distinct selection regimes that involve factors such as limited resources, reduced interspecific competition and a paucity of predators and parasites[6,7,9–13]. Although there are numerous examples of insular effects on mammals of the Cenozoic era[6–12,14–17], the effects of long-term isolation on islands are virtually unknown among Mesozoic mammaliaforms, and Mesozoic biotas more generally. Here we describe and analyse a complete, well-preserved skeleton of a gondwanatherian mammal (Fig. 1, Extended Data Fig. 1) from the latest Cretaceous period of Madagascar, which was then—and still remains—an island. This skeleton reveals an array of unusual and even unique adaptations that we hypothesize are due to evolution in an insular environment.

The new specimen—designated University of Antananarivo (UA) 9030—is the holotype of a new genus and species of gondwanatherian, *Adalatherium hui*, which we assign to the new family Adalatheriidae. UA 9030 is so well preserved that the distalmost caudal vertebrae, tiny phalangeal sesamoids and even non-osseous tissues (such as costal cartilage) are preserved in articular relationships. *A. hui* has an estimated body mass of about 3.1 kg (Extended Data Fig. 2, Supplementary Information, Supplementary Table 1). As such, it is the third-largest known mammal represented by anything more than isolated jaws and teeth from the Mesozoic era of Gondwana, despite the fact that it is represented by an immature individual.

Gondwanatheria, a clade restricted to Late Cretaceous and Palaeogene horizons of Gondwana, is particularly poorly represented in the fossil record[4,5]. Prior to the discovery reported here, the cranium of *Vintana*—also from the latest Cretaceous period of Madagascar—was the only gondwanatherian represented by more than isolated dental or gnathic remains[3,4]. The new fossil greatly expands our knowledge of gondwanatherians and of Mesozoic mammaliaforms from Gondwana in general, the fossil record of which is extremely limited relative to that of Laurasia[1,2].

Mammalia, Linnaeus 1758
Allotheria, Marsh 1880
Gondwanatheria, Mones 1987

[1]Department of Earth Sciences, Denver Museum of Nature and Science, Denver, CO, USA. [2]Department of Anatomical Sciences, Stony Brook University, Stony Brook, NY, USA. [3]Department of Anatomy, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY, USA. [4]Section of Mammals, Carnegie Museum of Natural History, Pittsburgh, PA, USA. [5]Department of Anatomical Sciences and Neurobiology, University of Louisville, Louisville, KY, USA. [6]Department of Anthropology, University of Texas at Austin, Austin, TX, USA. [7]Jackson School Museum of Earth History, University of Texas at Austin, Austin, TX, USA. [8]Department of Biomedical Sciences, Heritage College of Osteopathic Medicine, Ohio University, Athens, OH, USA. [9]Geology Department, Macalester College, St Paul, MN, USA. [10]Department of Anthropology, Stony Brook University, Stony Brook, NY, USA. [11]Institut für Geowissenschaften, Universität Bonn, Bonn, Germany. [12]School of Biological Sciences, Monash University, Clayton, Victoria, Australia. [13]Geosciences, Museums Victoria, Melbourne, Victoria, Australia. [14]Département de Sciences de la Terre et de l'Environnement, Université d'Antananarivo, Antananarivo, Madagascar. [15]Deceased: Yaoming Hu. ✉e-mail: David.Krause@dmns.org

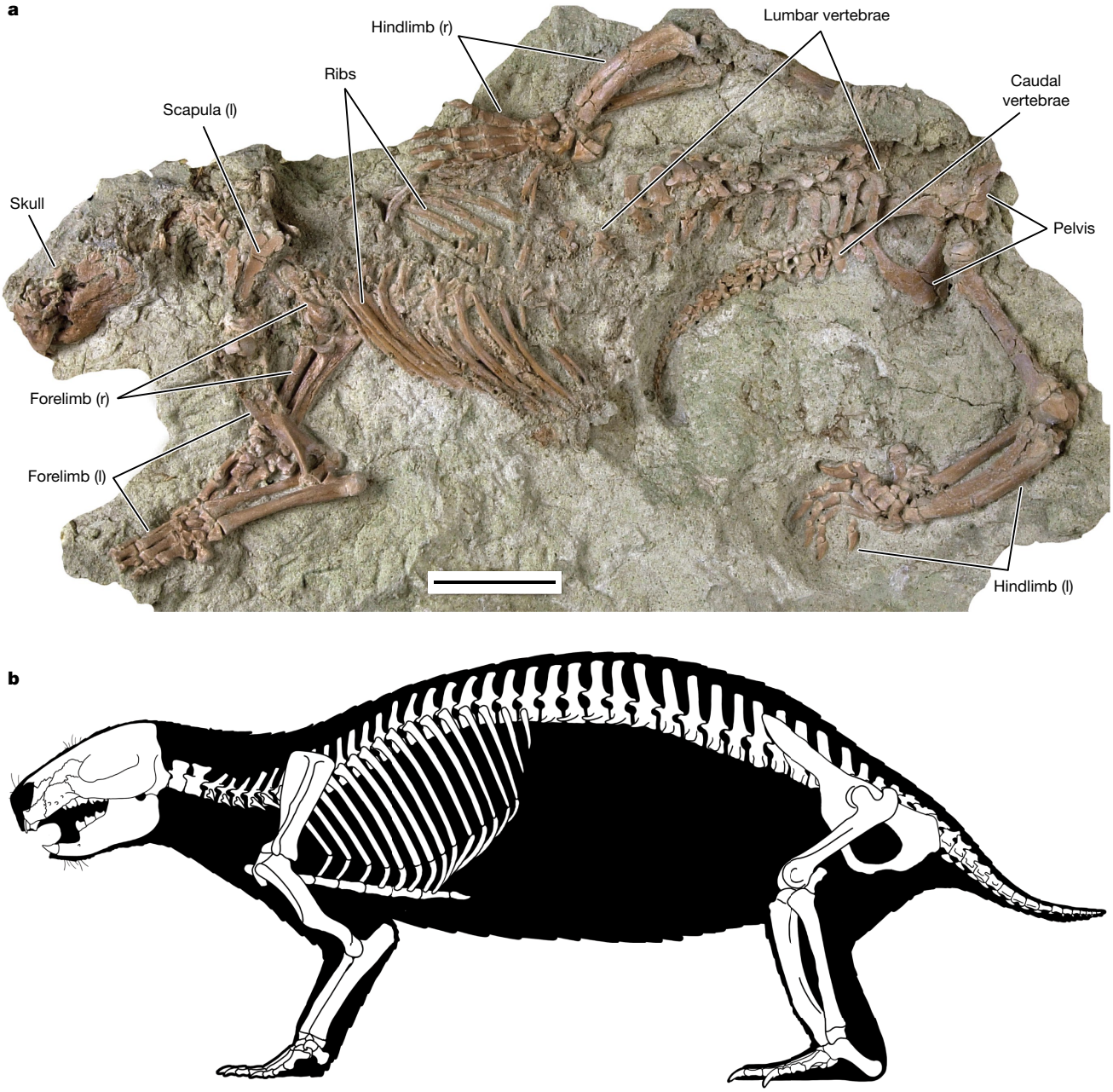


Fig. 1 | Skull and postcranial skeleton of *A. hui* holotype (UA 9030). a, 'Top' view, as preserved. Scale bar, 5 cm. **b**, Reconstruction in left lateral view. Left and right sides indicated as (l) and (r), respectively.

Adalatheriidae Krause, Hoffmann, Wible, and Rougier, 2020, fam. nov.
Adalatherium hui Krause, Hoffmann, Wible, and Rougier, 2020, gen. et sp. nov.

Etymology. *Adàla* (Malagasy), 'crazy'; *therium* (Latinized form of the Greek *θηριον*), 'beast'; the species name *hui* is in reference to the late Yaoming Hu for his contributions to our knowledge of early mammals.
Holotype. UA 9030, skull and postcranial skeleton.
Type locality and horizon. MAD99-15, Berivotra study area (northwestern Madagascar). Upper Cretaceous series (Maastrichtian stage, 72.1–66 million years (Myr) ago), Anembalemba Member, Maevarano Formation, Mahajanga Basin[18]. Additional information on the geological context is provided in the Supplementary Information.
Diagnosis. *A. hui* differs from all other known Mesozoic mammaliaforms in possessing quadrangular upper postcanine tooth crowns with four major cusps and three connecting perimetric ridges mesially, lingually and distally that border—on three sides—a central valley that opens buccally; and lower postcanine tooth crowns with four major cusps arranged in a diamond pattern and connected by four perimetric crests, and a prominent mesiobuccal basin on the two distalmost lower postcanines. The full diagnosis is provided in the Supplementary Information.

Cranium

The cranium of *Adalatherium* reveals a mosaic of plesiomorphic and derived features (Fig. 2a–d, Extended Data Fig. 3, Supplementary Videos 1–3). The presence of a very large internasal vacuity, five infraorbital foramina, a large foramen in the lacrimal that is not related to the nasolacrimal duct (probably for the ethmoidal branch of ophthalmic nerve (V_1) and associated vessels), numerous nasal foramina and a paranasal sinus that arises from the anterior vestibule of the nasal

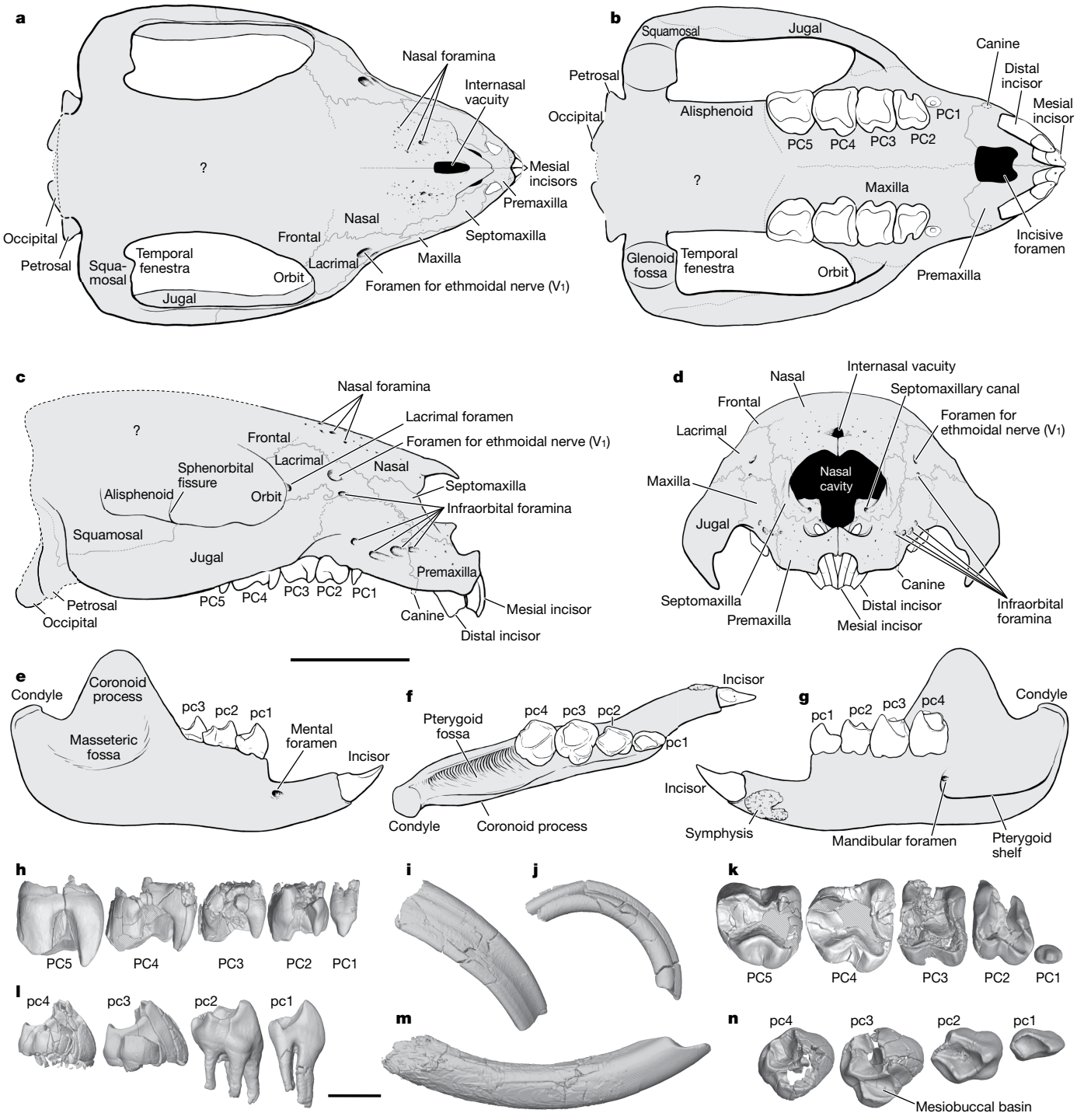


Fig. 2 | Cranium, lower jaw and dentition of *A. hui* holotype (UA 9030). **a**–**d**, Reconstructed cranium in dorsal (**a**), ventral (**b**), right lateral (**c**) and anterior (**d**) views. **e**–**g**, Reconstructed right lower jaw in lateral (**e**), dorsal (= occlusal) (**f**) and medial (**g**) views. **h**–**k**, Micro-computed tomography (μCT) digital renderings of right upper dentition, showing the postcanine teeth (**h**), distal incisor (**i**) and mesial incisor (**j**) in buccal views, and the postcanine teeth in occlusal view (**k**). **l**–**n**, μCT digital renderings of right lower dentition, showing the postcanine teeth (**l**) and incisor (**m**) in buccal views, and the postcanine teeth in occlusal view (**n**). Scale bars, 2 cm (**a**–**g**; scale bar above **e** and **f** applies to **a**–**g**), 5 mm (**h**–**n**). PC, upper postcanine tooth; pc, lower postcanine tooth.

cavity are particularly unusual for mammaliaforms. The snout region shares several features with that of *Vintana*[3,4], including the presence of a massive lacrimal bone that excludes the frontal from contacting the maxilla and a large septomaxilla with prominent posterodorsal and intranarial processes. By contrast, *Adalatherium* does not possess several autapomorphic features that are seen in *Vintana*, including a massive jugal flange and contact between the premaxillae and palatines.

Although much of the posterior part of the cranium was severely damaged post mortem, the left inner ear of UA 9030 is partially preserved and exhibits several features that were previously unknown

among mammaliaforms (Extended Data Fig. 4). Most notably, the primary bony lamina is structurally different from that of therians in being single-layered instead of double-layered, and the branches of the cochlear nerve appear to have passed along the surface of—rather than within—the primary lamina. This unique morphology suggests that the primary bony lamina of *Adalatherium* evolved convergently with those of therian mammals. The cochleae of *Adalatherium* and *Vintana* are unique among mammaliaforms in possessing a secondary bony canal that parallels the cochlear ganglion canal, and probably enclosed a vascular network. Overall, the cochlear canal is curved through at least 210° and possesses, in addition to the primary bony lamina, the base of a secondary bony lamina, a cribriform plate, a well-developed cochlear ganglion canal and a separate canal for the lagenar nerve; this last feature is not present in *Vintana*[4].

Lower jaw

The lower jaw of *Adalatherium* is more complete than in any other known gondwanatherian, and is the first to preserve the ascending ramus of the dentary (Fig. 2e–g, Extended Data Fig. 5, Supplementary Videos 4–7). Among gondwanatherians, the horizontal ramus of *Adalatherium* is essentially identical to that of *Sudamerica*[19] but differs from that of *Galulatherium*[5], primarily in having a stepped differential in height between the diastema and the postcanine alveolar portion. The dentary of *Adalatherium* is short and deep, and bears a large diastema between the incisor and postcanine teeth, a prominent pterygoid fossa and shelf, and a masseteric fossa that extends anteriorly onto the horizontal ramus. There is no evidence of a postdentary trough, Meckelian sulcus, coronoid bone or angular process. In these features, the dentary of *Adalatherium* is similar to those found in members of Multituberculata, which is a largely Laurasian group. The dentaries of euharamiyidans differ from that of *Adalatherium* in possessing an angular process, a coronoid bone and—according to ref. [20]—a 'reduced' postdentary trough (although this is disputed in ref. [21]). The dentary of *Haramiyavia* is much more plesiomorphic than those of both *Adalatherium* and euharamiyidans in retaining a long and shallow horizontal ramus, a fully developed postdentary trough and Meckelian sulcus, and in lacking a pterygoid fossa and shelf[22]. The masseteric fossa of *Adalatherium* is positioned relatively high dorsally on the dentary, an apparent autapomorphy.

Dentition

The dentition of *Adalatherium* is unlike that of any known Mesozoic mammaliaform (Fig. 2h–n, Supplementary Video 8 for upper postcanines). There are two very large, open-rooted upper incisors, each of which bears a buccally restricted band of enamel. The size, shape and positional relationships of the upper incisors are very similar to those discerned from the alveoli of *Vintana*[3,4]. The presence of upper canines in *Adalatherium* is indicated by tiny elliptical alveoli that are separated mesially from the incisors and distally from the postcanines by sizeable diastemata. The first upper postcanine is a small, simple, two-rooted premolariform tooth. The four more-distal postcanines are each supported by five or more roots and are unique among Mesozoic mammaliaforms in bearing four major cusps connected by ridges mesially, lingually and distally that border—on three sides—a central valley that opens buccally.

The single lower incisor is large, curved and open-rooted, and bears enamel that is largely restricted to the buccal surface. In these features, the lower incisor resembles those known for other gondwanatherians (except for the enamel-less condition in *Galulatherium*[5]). Each of the four lower postcanines has four cusps connected by prominent crests, forming a diamond pattern. The most mesial cusp dominates the crown on all four teeth. The second lower postcanine bears a mesiobuccal bulge, which is developed into a prominent basin on the penultimate and ultimate lower postcanines. The first lower postcanine has two roots, whereas each of the more distal postcanines has at least four.

The enamel microstructure of *Aadalatherium* consists of relatively plesiomorphic 'normal' radial enamel (Extended Data Fig. 6), typical of several gondwanatherians from the Late Cretaceous epoch and Palaeogene subperiod of Argentina[3,4]. It is unlike the 'modified' radial enamel, with pronounced inter-row sheets of interprismatic matrix, documented for other gondwanatherians (*Lavanify* and *Vintana*) from the Late Cretaceous epoch of India and Madagascar[3,4].

Postcranial skeleton

UA 9030 includes the only postcranial material assigned to a gondwanatherian, and *Adalatherium* is only the fourth Mesozoic mammaliaform from Gondwana represented by articulated postcranial remains[1]. The postcranial skeleton exhibits a number of unusual features, including an anteroposteriorly bowed and mediolaterally compressed tibia, a trochleated surface on the distal astragalus, a large number of trunk vertebrae (at least 16 thoracic and 12 lumbar vertebrae), and a short tail (24 vertebrae, almost all wider than they are long) (Figs. 1, 3, Extended Data Figs. 1, 7, 8). The long spinous and transverse processes of the thoracic and lumbar vertebrae suggest the presence of enhanced epaxial (back) musculature. In the pectoral girdle, a procoracoid bone is absent but a separate coracoid is well-developed (Fig. 3a, Extended Data Fig. 9a, b). The forelimbs had a moderately parasagittal posture, as indicated by the ventrally directed glenoid fossa and the well-developed humeral trochlea (Fig. 3a, d, Extended Data Fig. 8a, b). By contrast, the asymmetrical medial and lateral condyles of the femur are suggestive of a more sprawled hindlimb posture. Other notable features of the pelvis and hindlimbs include the presence of a large obturator foramen (similar in size to those of therians), an epipubic bone and a large, separate parafibula (Fig. 3, Extended Data Fig. 9c, d).

Phylogenetic relationships

Our phylogenetic analysis, performed using 84 cynodont taxa and 530 morphological characters, places *Adalatherium* within Gondwanatheria, which in turn is placed within Allotheria as the sister taxon to Multituberculata (Extended Data Fig. 10, Supplementary Information). Adalatheriidae (as solely represented by *Adalatherium*) is recovered as more derived than Ferugliotheriidae and stemward relative to Sudamericidae.

Previous phylogenetic analyses that include the recently discovered Early Cretaceous purported haramiyidan *Cifelliodon*[23,24] advanced the idea that gondwanatherians are nested within Eleutherodontida, basal to the purported Early Cretaceous hahnodontids *Hahnodon*[25] and *Cifelliodon*[23]. Although *Hahnodon* was not included in our analysis because it is represented by only one (or possibly two) isolated teeth[25,26], *Cifelliodon* is recovered at the base of Allotheria, which also includes Euharamiyida and 'Multituberculata + Gondwanatheria'. Our analysis places the haramiyidans *Haramiyavia* and *Thomasia*—along with the poorly known taxon *Megaconus*—outside of Mammaliaformes, with no close relations to allotherians. This finding is in contrast to previous analyses for *Vintana*[3] and for Jurassic euharamiyidans[27–29].

Evolution in isolation

Among mammals, the most obvious and quantifiable influences of evolving on islands are those related to body size. This observation has led to the articulation of the 'island rule', which states that—evolutionarily—small mammals on islands increase, and large mammals decrease, in size[11,30]. In addition, evolution in insular environments is thought to result in changes in anatomy, physiology, behaviour and life-history strategies, and (at the faunal level) relatively low species richness, taxonomic imbalance, high endemism and a general level of 'primitiveness'[6,12]. Although somewhat controversial and clearly not ubiquitous[31], the island rule is

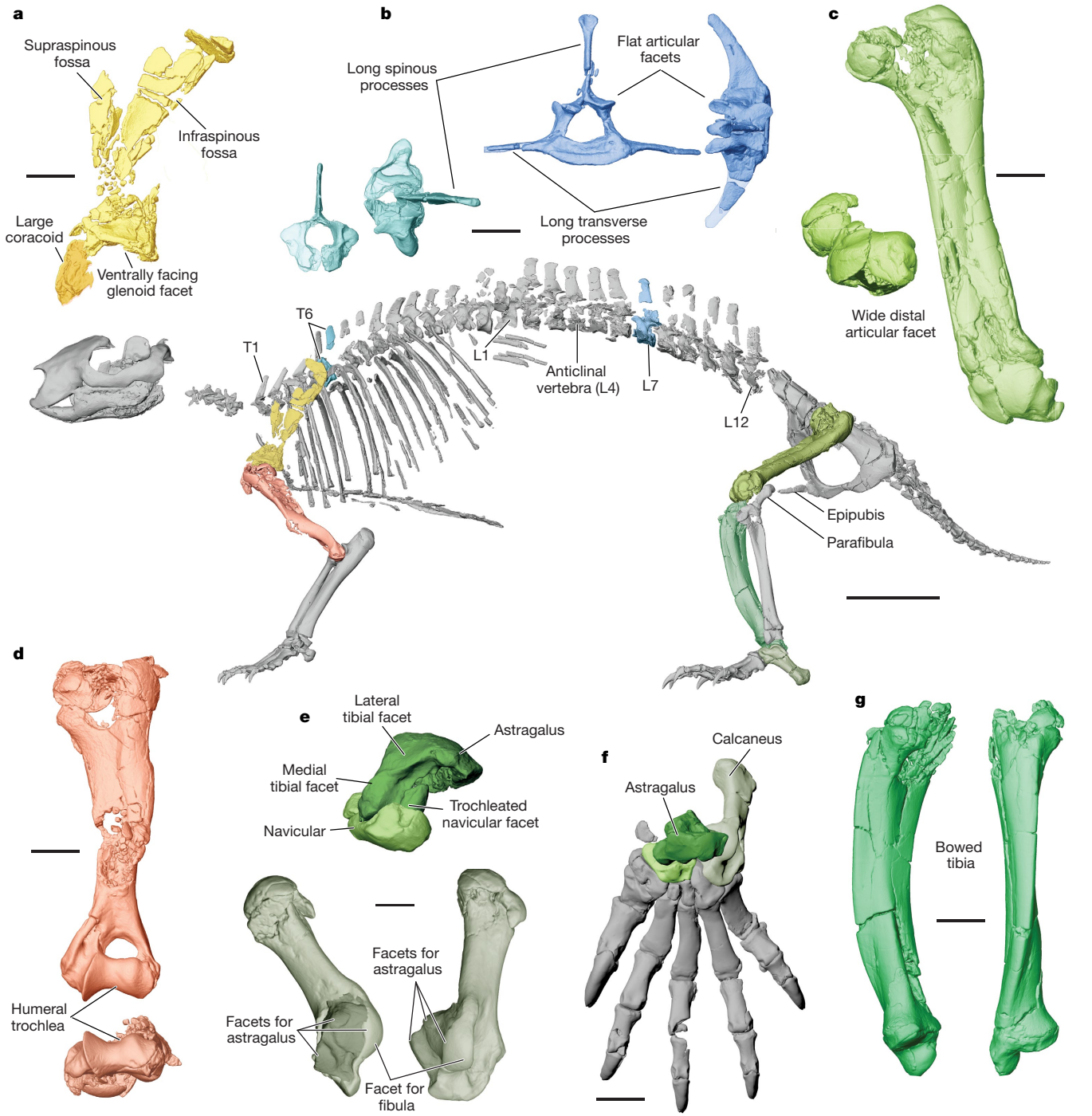


Fig. 3 | Skeleton of *A. hui* holotype (UA 9030). a–g, Digitally reconstructed skeleton in left lateral view, highlighting the left scapulocoracoid in lateral view (**a**), thoracic vertebra 6 and lumbar vertebra 7 in anterior and dorsal views (missing parts mirrored and rendered as semi-transparent) (**b**), the left femur in distal and anterior views (**c**), the left humerus in anterior and distal views (**d**), the left astragalus and navicular in anterior view, and left calcaneus in medial and dorsal views (**e**), left hind foot in dorsal view (**f**), and the left tibia in lateral and anterior views (**g**). L, lumbar vertebra; T, thoracic vertebra. Scale bars, 5 cm (main skeleton), 1 cm (**a**–**d**, **f**, **g**), 5 mm (**e**).

generally established as a pervasive pattern[9,11,32]. Examples of insular 'dwarfism' and 'gigantism' from Pliocene, Pleistocene and Holocene epochs abound[10,11,15], including from Madagascar (pygmy hippopotamuses[33] and giant lemurs[34]). Examples from earlier in the Cenozoic era are relatively sparse[14,16,17] and the effects of long-term isolation are extremely poorly documented for Mesozoic mammaliaforms.

Among Mesozoic mammaliaforms, adaptations related to evolution in isolation have been most notably claimed for two island environments, both from the latest part of the Cretaceous period (Maastrichtian age): (1) the gondwanatherian *Vintana* from Madagascar[3] and (2) the multituberculates *Barbatodon* and *Litovoi* from the archaic 'Haţeg Island'[35,36] (now part of Romania). Whether *Barbatodon* and *Litovoi* were part of a fauna that developed unique adaptations attributable to evolution in an insular environment is questionable—as is whether Haţeg Island was even an island (Supplementary Information). Because of its completeness and undoubted existence in an insular environment, the skeleton of *Adalatherium* provides an opportunity to examine evolution in isolation among Mesozoic mammaliaforms.

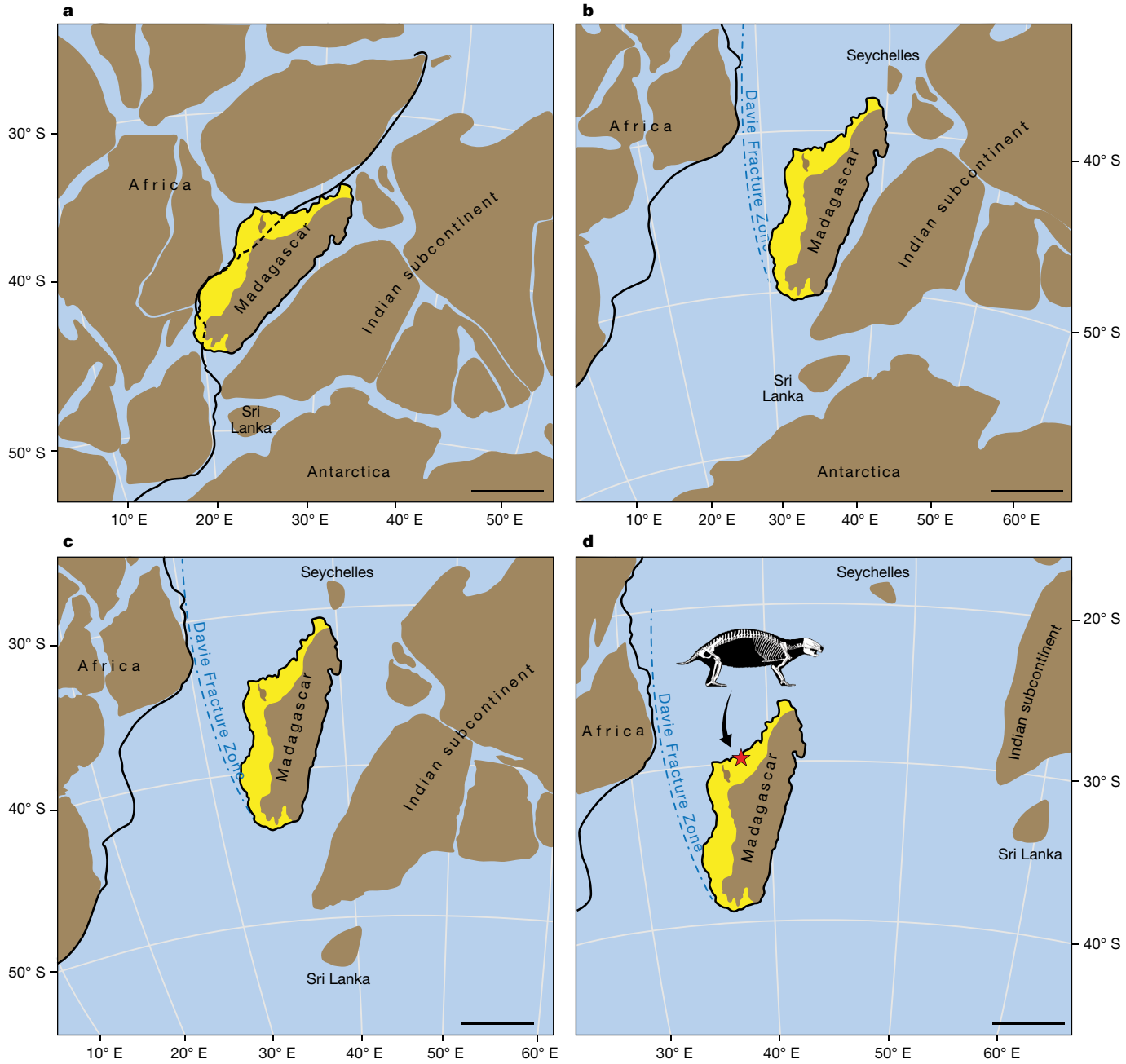


Fig. 4 | Key stages in plate tectonic history of Madagascar. a, Position of Madagascar before rifting between West Gondwana (South America and Africa) and East Gondwana (Madagascar, Seychelles, Indian subcontinent, Sri Lanka, Antarctica and Australia) at 183 Myr ago (Early Jurassic epoch). **b**, Separation of Indo-Madagascar from Antarctica and Australia at 124 Myr ago (mid-Early Cretaceous epoch). **c**, Separation of Indian subcontinent from Madagascar at 88 Myr ago (mid-Late Cretaceous epoch). **d**, Approximate time of deposition of Maevarano Formation at 66 Myr ago (latest Cretaceous period). Solid black lines indicate current coastlines of Madagascar and east Africa; brown represents Precambrian terranes; and yellow indicates sedimentary basins along west coast of Madagascar. The discovery site of UA 9030 is indicated by red star in **d**. Scale bars, 500 km. Maps adapted from Earthworks (www.reeves.nl/gond.com).

Madagascar separated from the Indian subcontinent and the Seychelles about 88 Myr ago[37,38] (Fig. 4). As a result, after separation, the obligate terrestrial taxa of Madagascar evolved in complete isolation and the only taxa that gained access to the island subsequently were flying, swimming or rafting forms that were able to disperse across considerable marine barriers[38–41]. Madagascar remains today a large, isolated continental block that is topographically high, geotectonically stable and at a minimum distance of 430 km from the closest mainland (Africa).

The postcranial skeleton of UA 9030 indicates that *Adalatherium* (and perhaps other gondwanatherians) was neither volant nor aquatic; it was an obligate terrestrial form that was relatively less capable of dispersal across marine barriers and more likely to have evolved on Madagascar. There are at least two other gondwanatherians that lived on Madagascar contemporaneously with *Adalatherium*: *Lavanify* (based on two fragmentary isolated teeth[42] that are insufficient to be informative in the current context) and *Vintana* (based on a complete cranium[3,4]). The cranium and upper postcanine dentition of *Vintana* exhibit several features that are unknown among Mesozoic mammaliaforms, but the number of such features in *Vintana* are far fewer than those for *Adalatherium* (based on a complete skeleton). Furthermore, given its deeply nested phylogenetic position within Allotheria (Extended Data

Fig. 10), several additional features of *Adalatherium* clearly evolved convergently with those of non-allotherian mammals—particularly therians (Supplementary Information). Considered together, and in concert with the palaeogeographical history of Madagascar, we hypothesize that the unusual morphological attributes of these two gondwanatherians are due to long-term evolution in isolation in an island environment, paralleling the cases adduced for various mammals on Mediterranean islands during the Neogene subperiod[6–8,10]. Furthermore, both *Adalatherium* and *Vintana* are large, being among the largest mammaliaforms known from the entire Mesozoic era of Gondwana. *Vintana*, which is even larger than *Adalatherium*, is the second-largest known Mesozoic mammaliaform globally and the largest from Gondwana[3,4]. Although the fossil record of Mesozoic mammaliaforms (especially from Gondwana) is too poor to conclusively establish that the large size of *Vintana* and *Adalatherium* reflects island gigantism, such a conclusion is consistent with the island rule[6–13,30].

The vertebrate fauna associated with *Adalatherium* (Supplementary Table 2) also exhibits unique characteristics reflective of its relictual nature. For instance, relatively high numbers of derived features are seen in other terrestrial members of the latest Cretaceous fauna of Madagascar, including the ceratophryid frog *Beelzebufo*, the crocodyliform *Simosuchus* and the theropods *Majungasaurus* and *Masiakasaurus* (ref. [38] and references therein). None of these genera is known from any other landmass, which again attests to high endemicity (although the very poor Mesozoic fossil record from Gondwana must be acknowledged in this regard). Furthermore, many of these forms had ghost lineages that extend back to or before the Early–Late Cretaceous boundary (100 Myr ago)[43], thus suggesting that their ancestors had probably arrived on Indo-Madagascar before its separation from other Gondwanan landmasses (that is, via a deep-time vicariance event rather than overwater dispersal). Despite originating in different ways, we conclude that the latest Cretaceous insular vertebrate fauna of Madagascar was probably as unique relative to mainland faunas as it is today.

The currently known latest Cretaceous (Maastrichtian) vertebrate fauna of Madagascar, assembled as an island fauna over the course of more than 20 Myr (around 88–66 Myr ago), became completely or nearly completely extinct—presumably the result of the end-Cretaceous bolide impact and/or the penecontemporaneous nearby volcanic eruptions that resulted in the Deccan Traps of India[44]. Thereafter, assembly of a Madagascan fauna began afresh. The establishment of a new vertebrate fauna required the arrival of transoceanic dispersers, primarily from Africa, that encountered habitats that were largely to completely devoid of Maastrichtian antecedents[38–41].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; and details of author contributions and competing interests are available at https://doi.org/10.1038/s41586-020-2234-8.

1. Kielan-Jaworowska, Z., Cifelli, R. L. & Luo, Z.-X. *Mammals from the Age of Dinosaurs: Origins, Evolution, and Structure* (Columbia Univ. Press, 2004).
2. Meng, J. Mesozoic mammals of China: implications for phylogeny and early evolution of mammals. *Natl Sci. Rev.* **1**, 521–542 (2014).
3. Krause, D. W. et al. First cranial remains of a gondwanatherian mammal reveal remarkable mosaicism. *Nature* **515**, 512–517 (2014).
4. Krause, D. W. (ed.) *Vintana sertichi* (Mammalia, Gondwanatheria) from the Late Cretaceous of Madagascar. *J. Vertebr. Paleontol.* **34**, suppl. to issue 6 (2014).
5. O'Connor, P. M. et al. A new mammal from the Turonian–Campanian (Upper Cretaceous) Galula Formation, southwestern Tanzania. *Acta Palaeontol. Pol.* **64**, 65–84 (2019).
6. Sondaar, P. Y. in *Major Patterns in Vertebrate Evolution* (eds Hecht, M. K. et al.) 671–707 (Plenum, 1977).
7. Azzarolli, A. in *Palaeontology, Essential of Historical Geology* (ed. Gallitelli, E. M.) 193–213 (S.T.E.M. Mucchi, 1982).
8. Whittaker, R. J. & Fernández-Palacios, J. M. *Island Biogeography: Ecology, Evolution, and Conservation* (Oxford Univ. Press, 2007).
9. Losos, J. B. & Ricklefs, R. E. Adaptation and diversification on islands. *Nature* **457**, 830–836 (2009).
10. Van der Geer, A., Lyras, G., de Vos, J. & Dermitzakis, M. *Evolution of Island Mammals: Adaptation and Extinction of Placental Mammals on Islands* (Wiley-Blackwell, 2010).
11. Lomolino, M. V. Of mice and mammoths: generality and antiquity of the island rule. *J. Biogeogr.* **40**, 1427–1439 (2013).
12. Benton, M. J. et al. Dinosaurs and the island rule: the dwarfed dinosaurs from Haţeg Island. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **293**, 438–454 (2010).
13. McNab, B. K. Geographic and temporal correlations of mammalian size reconsidered: a resource rule. *Oecologia* **164**, 13–23 (2010).
14. Ünay, E., De Bruijn, H. & Saraç, G. The Oligocene rodent record of Anatolia: a review. *Deinsea* **10**, 531–537 (2003).
15. De Vos, J., van den Hoek Ostende, L. W. & van den Bergh, G. D. in *Biogeography, Time and Place: Distributions, Barriers and Islands* (ed. Renema, W.) 315–346 (Springer, 2007).
16. Worthy, T. H. et al. Miocene mammal reveals a Mesozoic ghost lineage on insular New Zealand, southwest Pacific. *Proc. Natl Acad. Sci. USA* **103**, 19419–19423 (2006).
17. Métais, G. et al. Eocene metatherians from Anatolia illuminate the assembly of an island fauna during deep time. *PLoS ONE* **13**, e0206181 (2018).
18. Rogers, R. R., Hartman, J. H. & Krause, D. W. Stratigraphic analysis of Upper Cretaceous rocks in the Mahajanga Basin, northwestern Madagascar: implications for ancient and modern faunas. *J. Geol.* **108**, 275–301 (2000).
19. Pascual, R., Goin, F. J., Krause, D. W., Ortiz-Jaureguizar, E. & Carlini, A. A. The first gnathic remains of *Sudamerica*: implications for gondwanathere relationships. *J. Vertebr. Paleontol.* **19**, 373–382 (1999).
20. Luo, Z.-X. et al. New evidence for mammaliaform ear evolution and feeding adaptation in a Jurassic ecosystem. *Nature* **548**, 326–329 (2017).
21. Mao, F.-Y. & Meng, J. A new haramiyidan mammal from the Jurassic Yanliao Biota and comparisons with other haramiyidans. *Zool. J. Linn. Soc.* **186**, 529–552 (2019).
22. Luo, Z. X., Gatesy, S. M., Jenkins, F. A. Jr, Amaral, W. W. & Shubin, N. H. Mandibular and dental characteristics of Late Triassic mammaliaform *Haramiyavia* and their ramifications for basal mammal evolution. *Proc. Natl Acad. Sci. USA* **112**, E7101–E7109 (2015).
23. Huttenlocker, A. K., Grossnickle, D. M., Kirkland, J. I., Schultz, J. A. & Luo, Z.-X. Late-surviving stem mammal links the lowermost Cretaceous of North America and Gondwana. *Nature* **558**, 108–112 (2018).
24. King, B. & Beck, R. Bayesian tip-dated phylogenetics: topological effects, stratigraphic fit and the early evolution of mammals. Preprint at bioRxiv https://doi.org/10.1101/533885 (2019).
25. Sigogneau-Russell, D. First evidence of Multituberculata (Mammalia) in the Mesozoic of Africa. *Neues Jahrb. Geol. Paläontol. Monatsh.* **2**, 119–125 (1991).
26. Butler, P. M. & Hooker, J. J. New teeth of allotherian mammals from the English Bathonian, including the earliest multituberculates. *Acta Palaeontol. Pol.* **50**, 185–207 (2005).
27. Han, G., Mao, F., Bi, S., Wang, Y. & Meng, J. A Jurassic gliding euharamiyidan mammal with an ear of five auditory bones. *Nature* **551**, 451–456 (2017).
28. Zheng, X., Bi, S., Wang, X. & Meng, J. A new arboreal haramiyid shows the diversity of crown mammals in the Jurassic period. *Nature* **500**, 199–202 (2013).
29. Bi, S., Wang, Y., Guan, J., Sheng, X. & Meng, J. Three new Jurassic euharamiyidan species reinforce early divergence of mammals. *Nature* **514**, 579–584 (2014).
30. Van Valen, L. Pattern and the balance of nature. *Evol. Theory* **1**, 31–49 (1973).
31. Meiri, S., Cooper, N. & Purvis, A. The island rule: made to be broken? *Proc. R. Soc. Lond. B* **275**, 141–148 (2008).
32. McClain, C. R., Durst, P. A. P., Boyer, A. G. & Francis, C. D. Unravelling the determinants of insular body size shifts. *Biol. Lett.* **9**, 20120989 (2013).
33. Weston, E. M. & Lister, A. M. Insular dwarfism in hippos and a model for brain size reduction in *Homo floresiensis*. *Nature* **459**, 85–88 (2009).
34. Jungers, W. L., Demes, B. & Godfrey, L. R. in *Elwyn Simons: A Search for Origins* (eds Fleagle, J. G. & Gilbert, C. C.) 343–360 (Springer, 2008).
35. Smith, T. & Codrea, V. Red iron-pigmented tooth enamel in a multituberculate mammal from the Late Cretaceous Transylvanian "Haţeg Island". *PLoS ONE* **10**, e0132550 (2015).
36. Csiki-Sava, Z., Vremir, M., Meng, J., Brusatte, S. L. & Norell, M. A. Dome-headed, small-brained island mammal from the Late Cretaceous of Romania. *Proc. Natl Acad. Sci. USA* **115**, 4857–4862 (2018).
37. Reeves, C. The position of Madagascar within Gondwana and its movements during Gondwana dispersal. *J. Afr. Earth Sci.* **94**, 45–57 (2014).
38. Krause, D. W., Sertich, J. J. W., O'Connor, P. M., Curry Rogers, K. & Rogers, R. R. The Mesozoic biogeographic history of Gondwanan terrestrial vertebrates: insights from Madagascar's fossil record. *Annu. Rev. Earth Planet. Sci.* **47**, 519–553 (2019).
39. Yoder, A. D. & Nowak, M. D. Has vicariance or dispersal been the predominant biogeographic force in Madagascar? Only time will tell. *Annu. Rev. Ecol. Evol. Syst.* **37**, 405–431 (2006).
40. Krause, D. W. Washed up in Madagascar. *Nature* **463**, 613–614 (2010).
41. Samonds, K. E. et al. Imperfect isolation: factors and filters shaping Madagascar's extant vertebrate fauna. *PLoS ONE* **8**, e62086 (2013).
42. Krause, D. W., Prasad, G. V. R., Koenigswald, W. V., Sahni, A. & Grine, F. E. Cosmopolitanism among Gondwanan Late Cretaceous mammals. *Nature* **390**, 504–507 (1997).
43. Ali, J. & Krause, D. W. Late Cretaceous bioconnections between Indo-Madagascar and Antarctica: refutation of the Gunnerus Ridge causeway hypothesis. *J. Biogeogr.* **38**, 1855–1872 (2011).
44. Schoene, B. et al. U–Pb constraints on pulsed eruption of the Deccan Traps across the end-Cretaceous mass extinction. *Science* **363**, 862–866 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The holotype of *A. hui* is catalogued into the University of Antananarivo collections. Graphics and phylogenetics data are provided in the Supplementary Information. The Life Science Identifiers (LSID) for the new family, genus and species are registered with Zoobank (http://zoobank.org) with the identifiers urn:lsid:zoobank.org:act:DA019E3D-6B27-4728-9321-A0502EA0FEC4, urn:lsid:zoobank.org:act:3DEB862B-5A3E-40A2-B939-A2B549A0AC62 and urn:lsid:zoobank.org:act:9CB87C13-10B6-4FE7-B450-BE89D4E6BD88, respectively. The data matrix for the phylogenetic analysis has been deposited in MorphoBank at http://morphobank.org/permalink/?P3411.

45. Campione, N. C. & Evans, D. C. A universal scaling relationship between body mass and proximal limb bone dimensions in quadrupedal terrestrial tetrapods. *BMC Biology* **10**, 60 (2012).

Acknowledgements We thank the Université d'Antananarivo, the Madagascar Institut pour la Conservation des Ecosystèmes Tropicaux and the villagers of the Berivotra Study Area for logistical support of fieldwork; the ministries of Mines and Higher Education of the Republic of Madagascar for permission to conduct field research; members of the 1999 field research team for their efforts; and the National Geographic Society and the National Science Foundation for funding. Full acknowledgments are provided in the Supplementary Information.

Author contributions D.W.K., S.H. and Y.H. conceived the project; J.R.G. and S.H. contributed to μCT digital preparation of the specimen; R.R.R. and L.J.R. provided geological data and interpretation; E.C.K., D.W.K. and S.H. developed the body mass estimates; D.W.K., S.H., Y.H., J.R.W., J.B.R, G.W.R., J.R.G., J.A.S., A.R.E., E.C.K. and W.v.K. conducted laboratory work on the fossil and contributed to descriptions and comparisons; S.H., D.W.K., J.R.W. and G.W.R. contributed to the phylogenetic analysis; D.W.K. wrote the manuscript, with contributions and/or editing from all authors.

Competing interests The authors declare no competing interests.

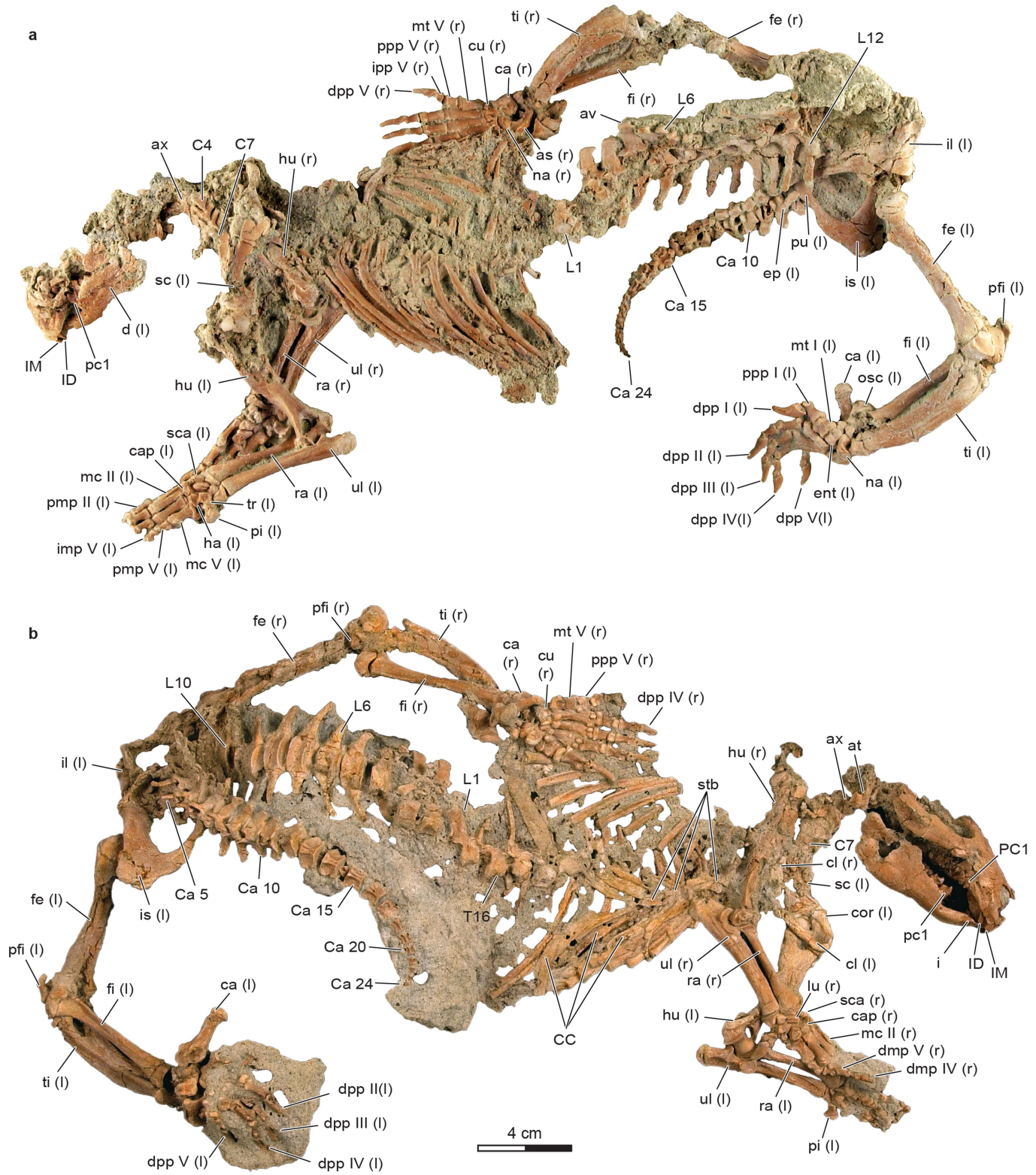
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2234-8.

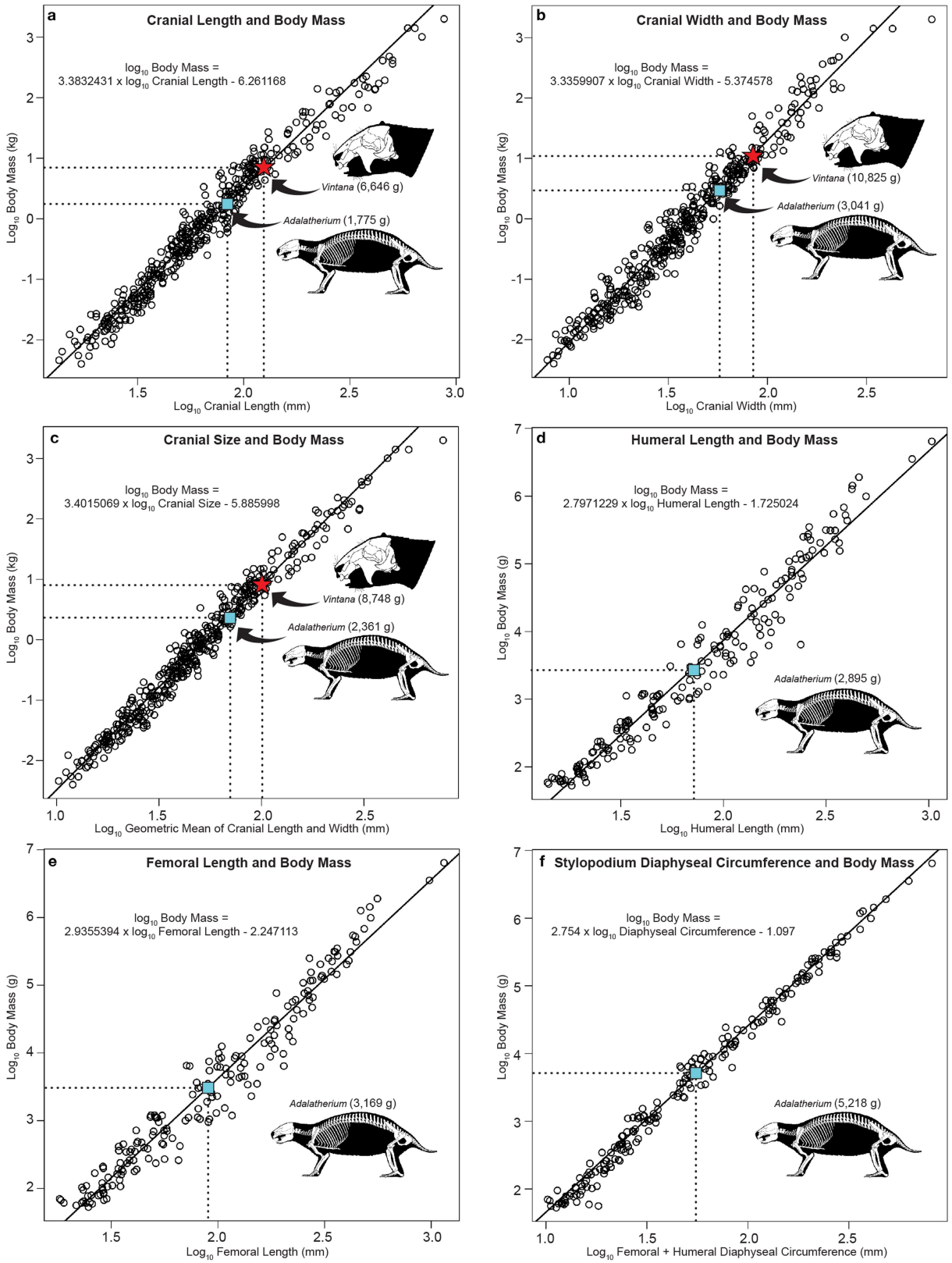
Correspondence and requests for materials should be addressed to D.W.K.

Peer review information *Nature* thanks Jin Meng and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

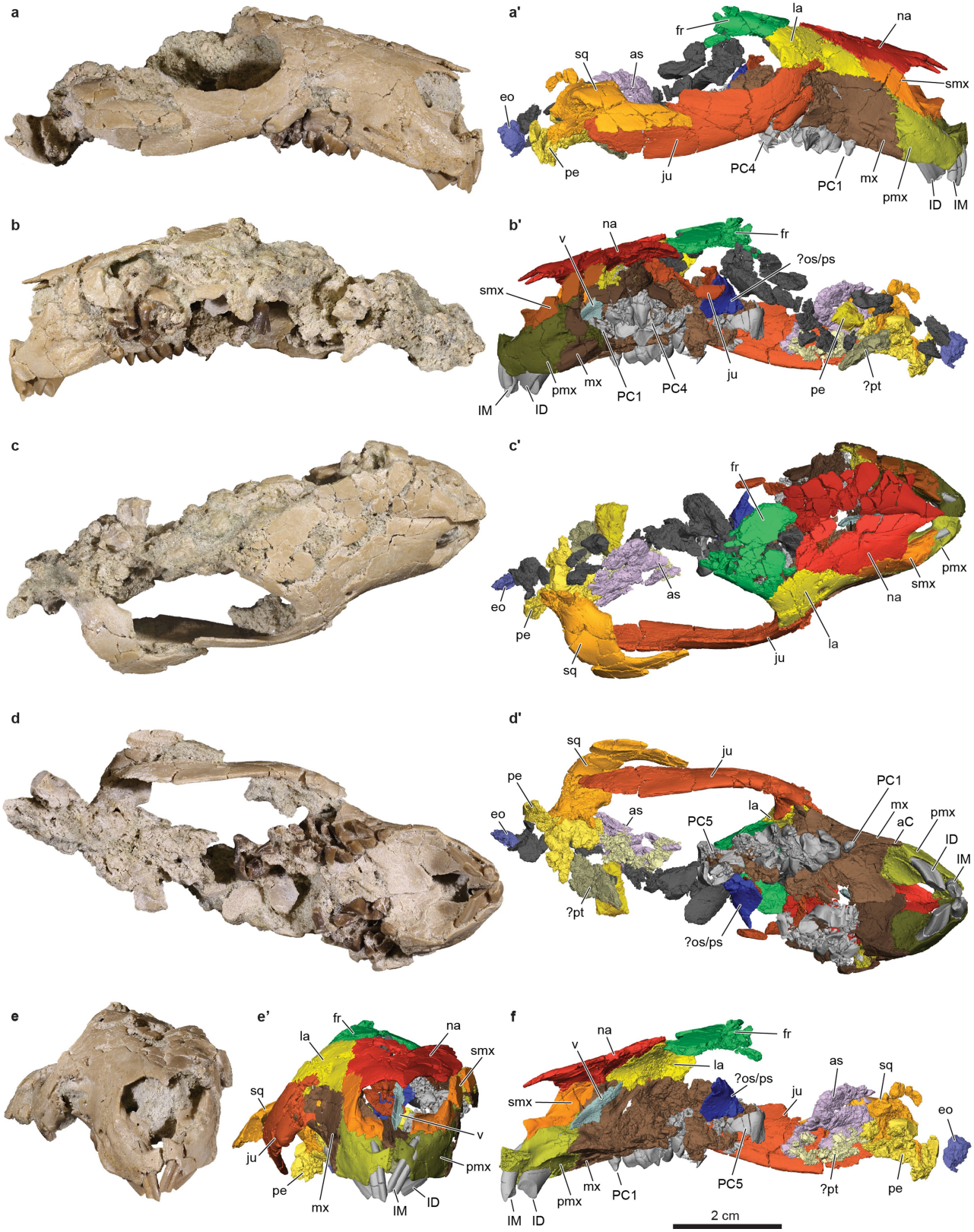
Reprints and permissions information is available at http://www.nature.com/reprints.



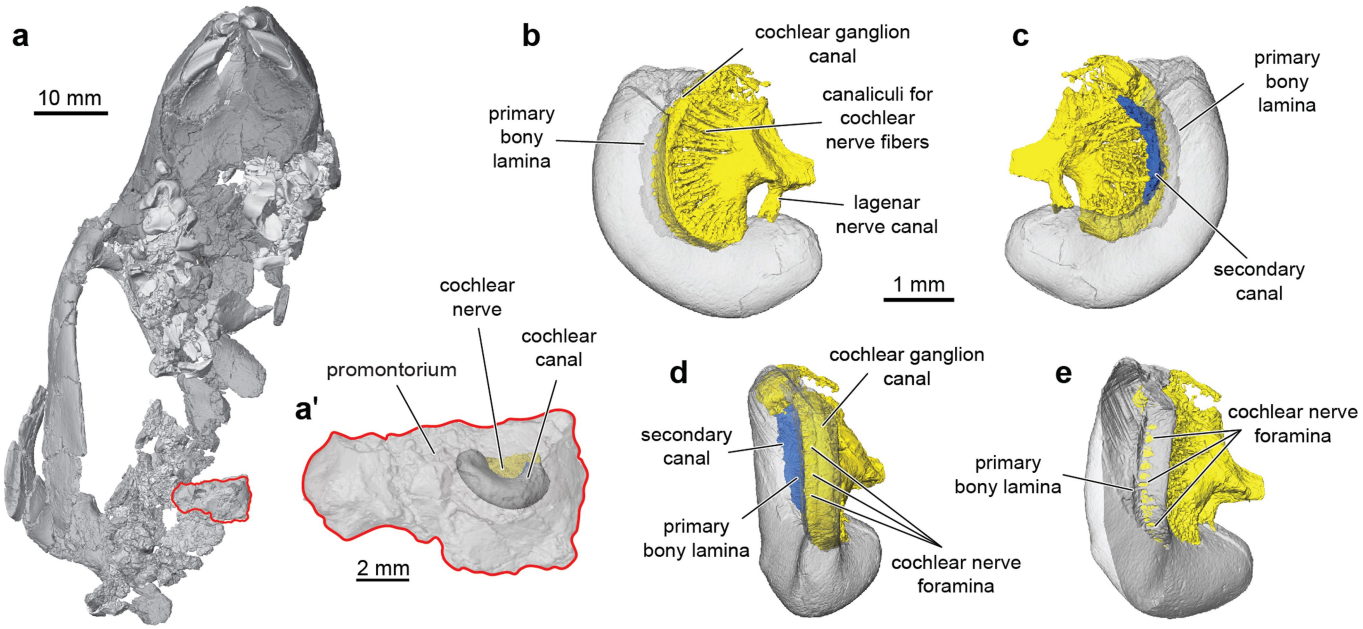
Extended Data Fig. 1 | Photographs of the skeleton of *A. hui* holotype (UA 9030). a, **b**, 'Top' (**a**) and 'bottom' (**b**) views, as preserved. The left and right sides are indicated as (l) and (r), respectively. as, astragalus; at, atlas; av, anticlinal vertebra; ax, axis; C, cervical vertebra; ca, calcaneus; Ca, caudal vertebra; cap, capitate; CC, costal cartilage; cl, clavicle; cor, coracoid; cu, cuboid; dpp, distal pedal phalanx; ent, entocuneiform; ep, epipubis; fe, femur; fi, fibula; ha, hamate; hu, humerus; i, lower incisor; ID, distal upper incisor; IM, medial upper incisor; imp, intermediate manual phalanx; ipp, intermediate pedal phalanx; L, lumbar vertebra; lu, lunate; m, mandible; mc, metacarpal; mt, metatarsal; na, navicular; osc, os calcaris; pc1, lower first postcanine tooth; PC1, upper first postcanine tooth; pe, pelvis; pfi, parafibula; pi, pisiform; pmp, proximal manual phalanx; ppp, proximal pedal phalanx; R, rib; ra, radius; sc, scapula; sca, scaphoid; stb, sternebra; T, thoracic vertebra; ti, tibia; tr, triquetrum; ul, ulna.



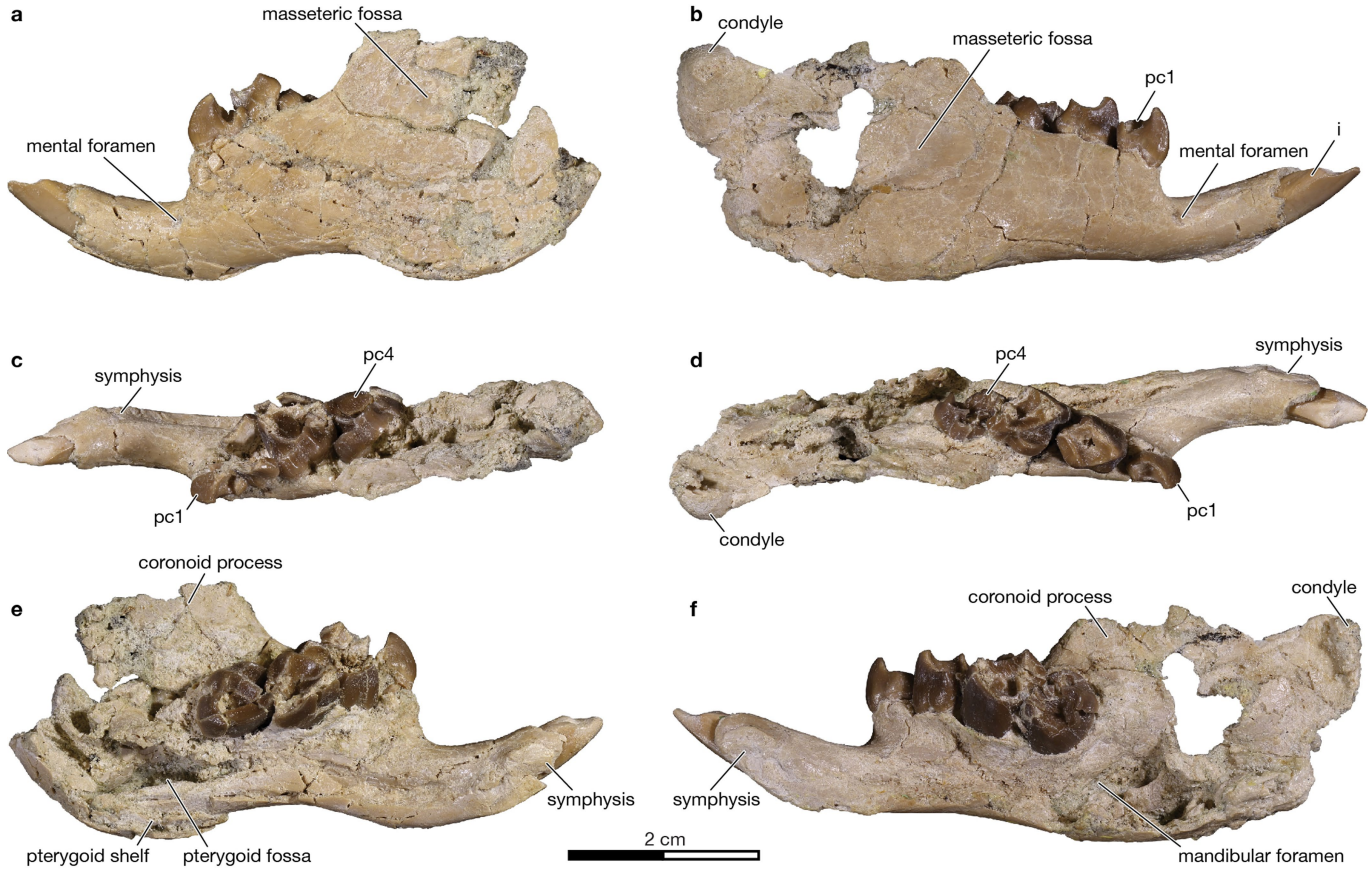
Extended Data Fig. 2 | Bivariate plots of body mass estimates for *A. hui*. **a**, Relationship between cranial length and body mass in 423 extant mammals, plus estimated body mass in the gondwanatherians *A. hui* and *Vintana sertichi*. **b**, Relationship between cranial width and body mass in 423 extant mammals, plus the estimated body mass in *A. hui* and *V. sertichi*. **c**, Relationship between cranial size and body mass in 423 extant mammals, plus the estimated body mass in *A. hui* and *V. sertichi*. **d**, Relationship between humeral length and body mass in 187 extant therian mammals, plus the estimated body mass in *A. hui*. **e**, Relationship between femoral length and body mass in 184 extant species of therian mammal, plus the estimated body mass in *A. hui*. **f**, Relationship between stylopodial diaphyseal circumference and body mass as calculated for a sample of 245 tetrapod species[45] (data points shown for mammals only, $n = 200$), plus the estimated body mass in *A. hui*. Regression lines in **a**–**e** are from ordinary least squares regressions, whereas the regression line shown in **f** is from a phylogenetic generalized least squares regression. Measurement data, methods and references are provided in the Supplementary Information.



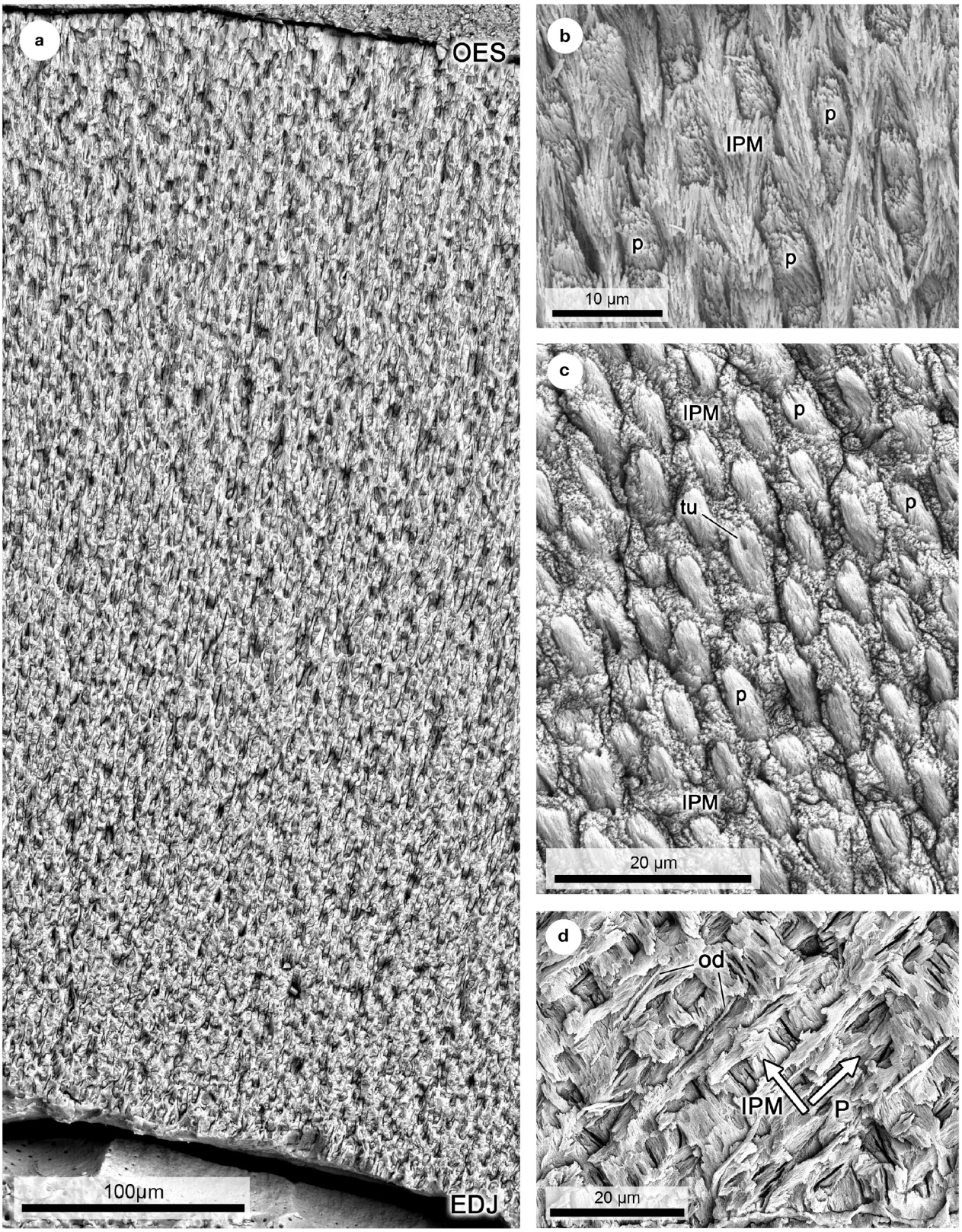
Extended Data Fig. 3 | Cranium of *A. hui* holotype (UA 9030). **a**–**e**, Photographs of external surfaces of cranium in right lateral (**a**), left lateral (**b**), dorsal (**c**), ventral (**d**) and anterior (**e**) views. **a'**–**e'**, Labelled μCT images of cranium in the same views as in **a**–**e**, respectively. **f**, Labelled μCT image of medial view of right side of nasal cavity. aC, alveolus for upper canine; as, alisphenoid; eo, exoccipital; fr, frontal; ID, distal upper incisor; IM, mesial upper incisor; ju, jugal; la, lacrimal; mx, maxilla; na, nasal; os/ps, orbitosphenoid/presphenoid complex; PC, upper postcanine tooth; pe, petrosal; pmx, premaxilla; pt, pterygoid; smx, septomaxilla; sq, squamosal; v, vomer.



Extended Data Fig. 4 | Inner ear of *A. hui* holotype (UA 9030). a, Ventral view of reconstructed cranium, with petrosal fragment bounded by red line enlarged in **a'**. **b**–**e**, Reconstructed cochlear canal in dorsomedial (**b**), ventrolateral (**c**) and posteroventromedial (**d**, **e**) views, with the view in **d** being slightly more posterior and the view in **e** slightly more medial. In **e**, only the medial aspect of cochlear canal in grey is shown, to reveal primary bony lamina and cochlear nerve foramina. Semi-transparent grey, cochlear canal; yellow, cochlear nerve; blue, secondary canal.

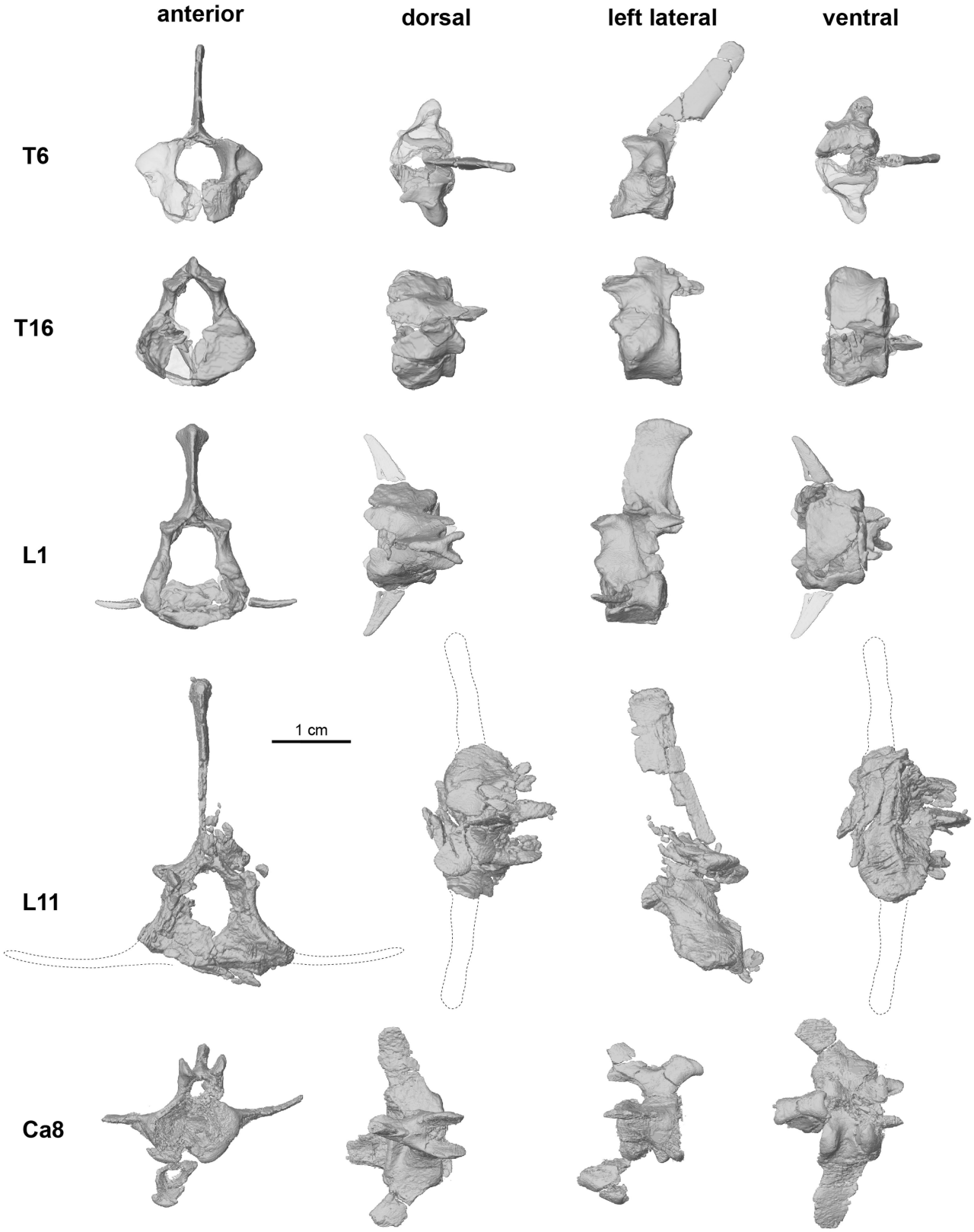


Extended Data Fig. 5 | Lower jaw of *A. hui* holotype (UA 9030). Photographs of left dentary in left column; photographs of right dentary in right column. **a**, **b**, Lateral views. **c**, **d**, Dorsal (occlusal) views. **e**, **f**, Medial views. i, lower incisor; pc, lower postcanine tooth.

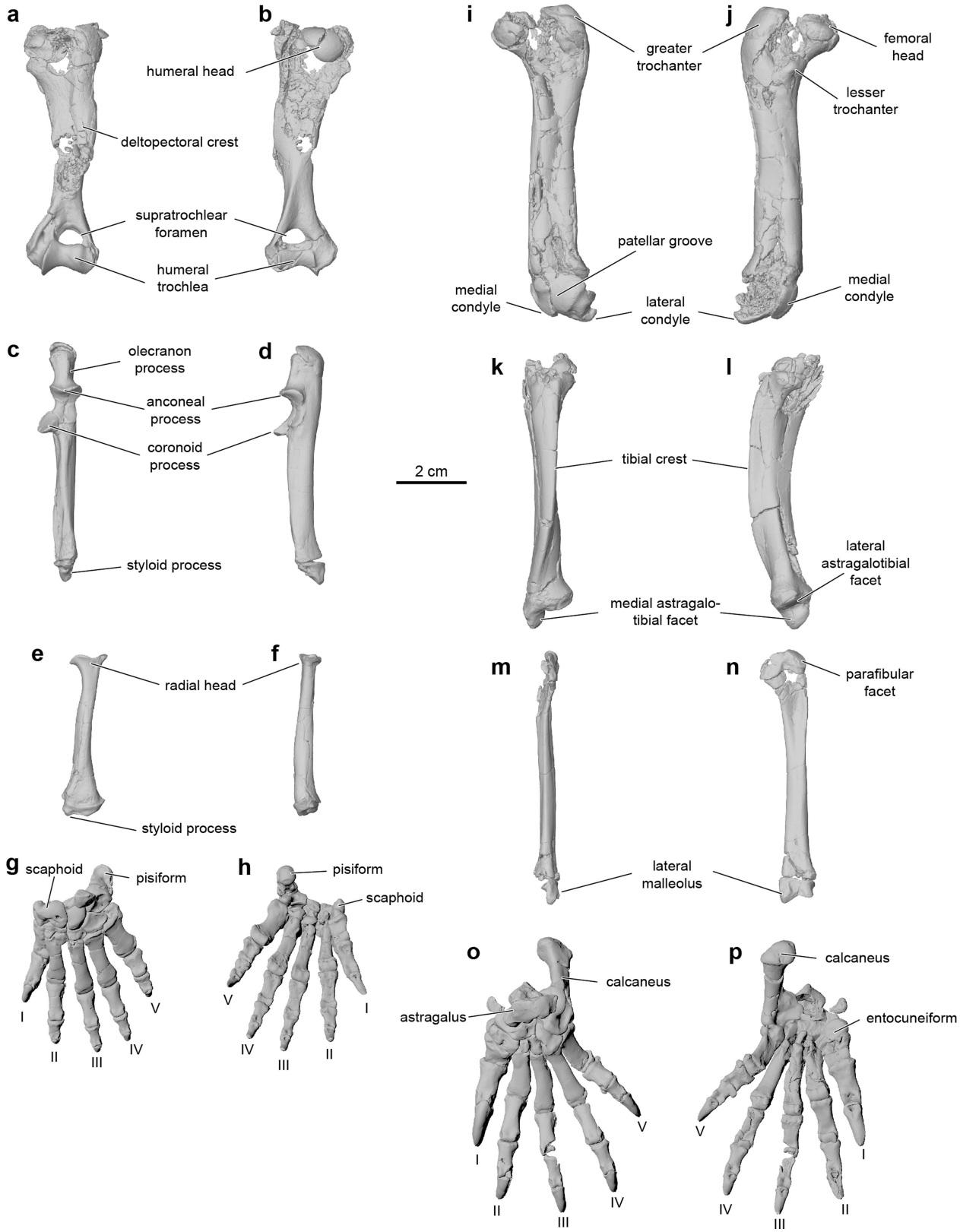


Extended Data Fig. 6 | See next page for caption.

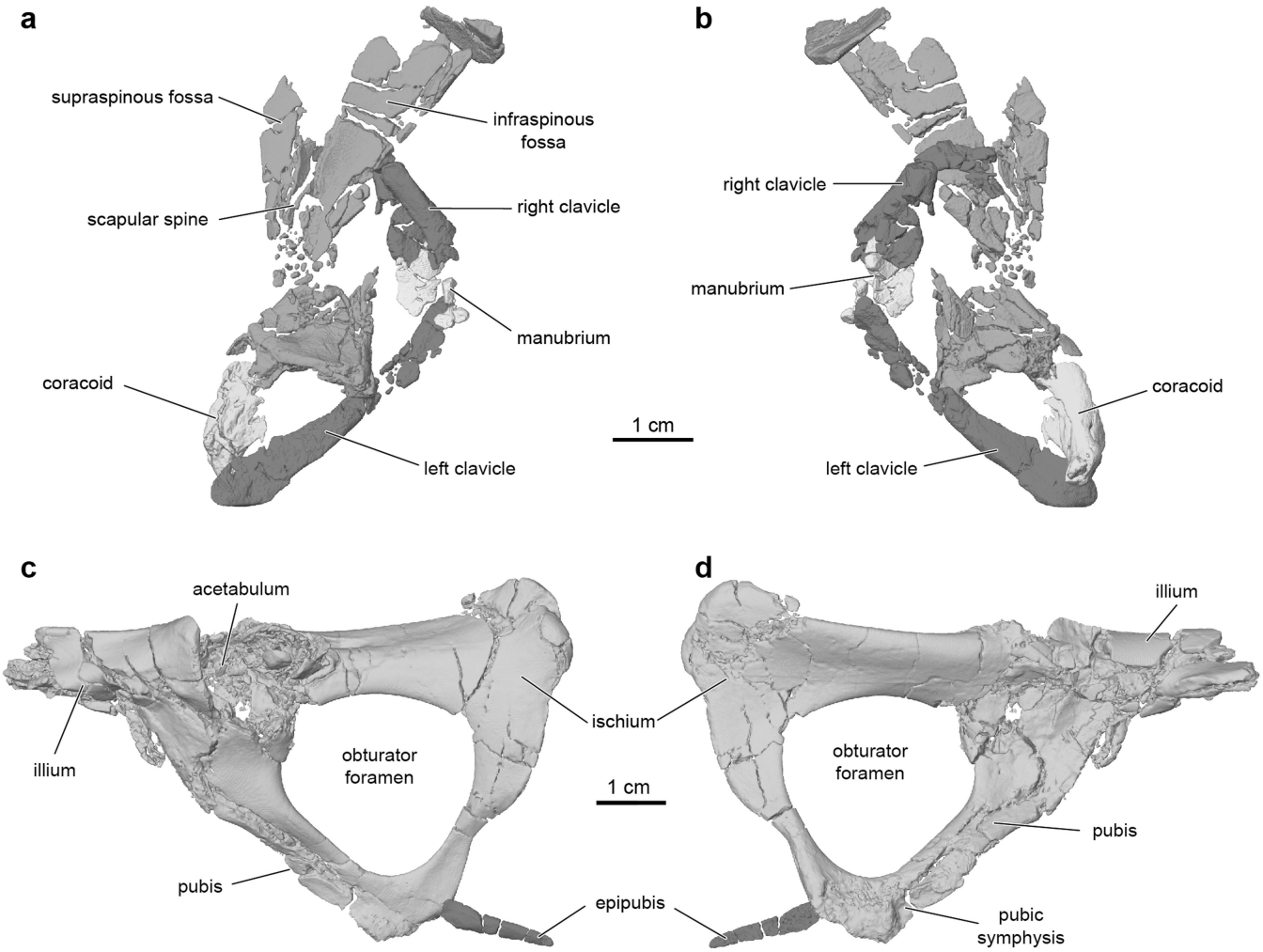
Extended Data Fig. 6 | Enamel microstructure of *A. hui* holotype (UA 9030). **a**–**d**, Scanning electron micrographs of single postcanine tooth enamel fragment sectioned in various planes. **a**, Transverse section of entire enamel band from the enamel–dentine junction (EDJ) to the outer enamel surface (OES) (about 0.4-mm thick) showing single layer of radial enamel and absence of distinct layer of prismless external enamel. Prism size increases from, on average, 2.3 to 2.8 µm from the enamel–dentine junction to the outer enamel surface. Prisms close to the enamel–dentine junction are intersected by interprismatic matrix at slightly higher angles than towards the outer enamel surface. **b**, Transverse section showing the clear distinction between enamel prisms and interprismatic matrix. **c**, Radial section showing radial enamel in outer zone with prisms surrounded by interprismatic matrix and some cross-sections of prisms showing tubules. **d**, Radial, but slightly oblique, section showing enamel of inner zone with prisms enveloped by interprismatic matrix and presence of odontoblastic processes. In this zone, crystallites of interprismatic matrix lie almost perpendicular to those of prisms. Prisms rise from the enamel–dentine junction at angle of about 45°; this angle is reduced only slightly towards the outer enamel surface. IPM, interprismatic matrix; od, odontoblastic process; p, prism; tu, tubule.



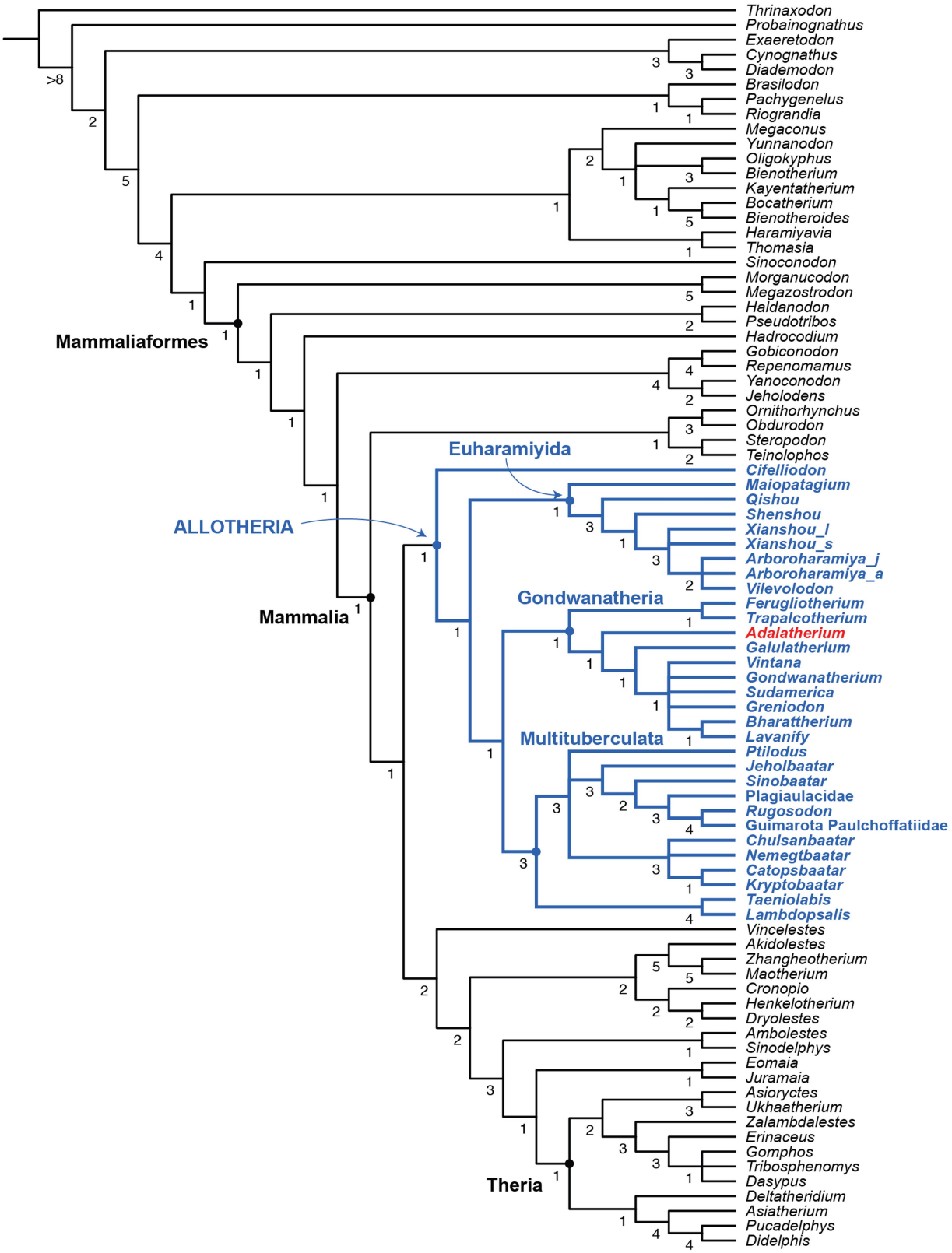
Extended Data Fig. 7 | Selected individual vertebrae of *A. hui* holotype (UA 9030). Thoracic (T6 and T16), lumbar (L1 and L11) and anterior caudal (Ca8) vertebrae are depicted in anterior, dorsal, left lateral and ventral views. The left transverse process of L11 is preserved but was separated from the vertebral column during preparation and has not been CT scanned. Dotted outlines represent the shape of preserved left transverse process, and the mirrored reconstructed right transverse process.



Extended Data Fig. 8 | Limb bone elements of *A. hui* holotype (UA 9030). **a**–**p**, μCT images. **a**, **b**, Left humerus in anterior (**a**) and posterior (**b**) views. **c**, **d**, Left ulna in anterior (**c**) and lateral (**d**) views. **e**, **f**, Left radius in anterior (**e**) and lateral (**f**) views. **g**, **h**, Left manus in dorsal (**g**) and palmar (=ventral) (**h**) views. **i**, **j**, Left femur in anterior (**i**) and posterior (**j**) views. **k**, **l**, Left tibia in anterior (**k**) and lateral (**l**) views. **m**, **n**, Left fibula in anterior (**m**) and lateral (**n**) views. **o**, **p**, Left pes in dorsal (**o**) and plantar (=ventral) (**p**) views.



Extended Data Fig. 9 | Pectoral and pelvic girdle elements of *A. hui* holotype (UA 9030). a–**d**, μCT images. **a**, **b**, Left scapulacoracoid, left and right clavicle and manubrium in 'top' (**a**) and 'bottom' (**b**) views (as preserved). **c**, **d**, Left os coxa and epipubic bone in lateral (**c**) and medial (**d**) views.



Extended Data Fig. 10 | Phylogenetic relationships of *A. hui* and selected mammaliaforms. Strict consensus tree of 16 equally parsimonious trees (tree length = 2,315, consistency index = 0.3015 and retention index = 0.7001) derived from analysis of 84 cynodont taxa and 530 characters, with multistate characters unordered and unweighted. Bremer values are listed next to the nodes. *Adalatherium* is highlighted in red. Allotheria—consisting of *Cifelliodon*, Euharamiyida, Gondwanatheria (including *Adalatherium*) and Multituberculata—is highlighted in blue. Taxon and character lists, the data matrix, limitations and assumptions, phylogenetic methods and a more detailed explanation of the results are provided in the Supplementary Information.

nature research

Corresponding author(s): David W. Krause

Last updated by author(s): Feb 15, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Avizo 7 (VSG), 8 (FEI), and 9 (FEI/Thermo-Fisher Scientific); Amira 6 (FEI/Thermo-Fisher Scientific - measurement tool, spline probe tool); Dragonfly 3.0; Animation Producer in Avizo; Adobe Premiere Pro (Creative Cloud edition); Geomagic Wrap (MeshDoctor and Relax functions); Autodesk 3Ds Max
Data analysis	TNT version 1.1; PAUP* 4.0; JAS version 14 (SAS Institute); Avizo 7 (VSG), 8 (FEI), and 9 (FEI/Thermo-Fisher Scientific); Amira 6 (FEI/Thermo-Fisher Scientific - measurement tool, spline probe tool); Dragonfly 3.0; Animation Producer in Avizo; Adobe Premiere Pro (Creative Cloud edition); Geomagic Wrap (MeshDoctor and Relax functions); Autodesk 3Ds Max

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The holotypic specimen of Adalatherium hui is reposited in the University of Antananarivo (UA), Madagascar. The data matrix for the phylogenetic analysis has been deposited in MorphoBank (http://morphobank.org/permalink/?P3411). The Life Science Identifiers (LSID) for the new family, genus, and species are registered with Zoobank (http://zoobank.org).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our study entails the description of a new fossil taxon represented by only a single specimen. Comparisons were made with many other taxa from around the world that are documented in detail in the Supplementary Information section. Comparisons with extant mammals were conducted in the large, diverse collections of the Department of Zoology, Denver Museum of Nature & Science and the Section of Mammals, Carnegie Museum of Natural History, Pittsburgh, USA.
Data exclusions	We sampled extinct taxa as broadly as possible, but were limited by the availability to study specimens of some taxa firsthand. In these instances, we used casts, 3D prints, CT scan datasets, and photographs.
Replication	Our taxon-character matrix is a derivative of that employed by Krause et al. (2014 – Nature) with improvements that are explicitly documented so as to allow replication.
Randomization	Our study only reports a single specimen; as such, no randomizations were possible.
Blinding	Character matrices for phylogenetic analysis were developed on the basis of independent observation of each taxon.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☐	☒ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Palaeontology

Specimen provenance	The holotypic specimen was collected from the Upper Cretaceous Maevarano Formation, Mahajanga Basin, Madagascar. The specimen was collected under a Collaborative Agreement with the University of Antananarivo and various ministries of the Madagascar government.
Specimen deposition	The University of Antananarivo (UA), Madagascar
Dating methods	No new dates were obtained for this study.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Olfactory sniffing signals consciousness in unresponsive patients with brain injuries

https://doi.org/10.1038/s41586-020-2245-5

Received: 25 September 2019

Accepted: 26 February 2020

Published online: 29 April 2020

Check for updates

Anat Arzi[1,2,3]✉, Liron Rozenkrantz[2,3], Lior Gorodisky[2,3], Danit Rozenkrantz[4], Yael Holtzman[2,3], Aharon Ravia[2,3], Tristan A. Bekinschtein[1], Tatyana Galperin[4], Ben-Zion Krimchansky[4], Gal Cohen[4], Anna Oksamitni[4], Elena Aidinoff[4], Yaron Sacher[4,5,6] & Noam Sobel[2,3,6]✉

After severe brain injury, it can be difficult to determine the state of consciousness of a patient, to determine whether the patient is unresponsive or perhaps minimally conscious[1], and to predict whether they will recover. These diagnoses and prognoses are crucial, as they determine therapeutic strategies such as pain management, and can underlie end-of-life decisions[2,3]. Nevertheless, there is an error rate of up to 40% in determining the state of consciousness in patients with brain injuries[4,5]. Olfaction relies on brain structures that are involved in the basic mechanisms of arousal[6], and we therefore hypothesized that it may serve as a biomarker for consciousness[7]. Here we use a non-verbal non-task-dependent measure known as the sniff response[8–11] to determine consciousness in patients with brain injuries. By measuring odorant-dependent sniffing, we gain a sensitive measure of olfactory function[10–15]. We measured the sniff response repeatedly over time in patients with severe brain injuries and found that sniff responses significantly discriminated between unresponsive and minimally conscious states at the group level. Notably, at the single-patient level, if an unresponsive patient had a sniff response, this assured future regaining of consciousness. In addition, olfactory sniff responses were associated with long-term survival rates. These results highlight the importance of olfaction in human brain function, and provide an accessible tool that signals consciousness and recovery in patients with brain injuries.

Sniff responses can be sensory-driven, cognitively driven, or both. Sensory-driven sniff responses reflect automatic odorant-driven modifications in nasal airflow, which are evident in humans[8,10,11] and other animals[9,16]. Sensory-driven sniff responses have two levels: level 1, odorant detection, which reflects the changes in nasal airflow in response to the presence of an odorant; and level 2, odorant discrimination, which is the differential response to different odorants, such as reduced nasal airflow when sensing an unpleasant compared with a pleasant odorant[8,12,13,17]. Cognitively driven sniff responses reflect situational understanding and/or learning. For example, patients in this study were told that they would be presented with odorants. If a patient then modifies their nasal airflow in response to an empty jar (used as the 'blank') when it was presented beneath their nose, this implies the possible awareness of the jar and/or possible learned anticipation of an odorant.

We used pleasant (shampoo) and unpleasant (rotten fish) odorants (which were valence-rated independently[12,13,18]) to trigger sensory sniff responses, and blank presentations to trigger cognitive sniff responses (Fig. 1a), in 43 patients with disorders of consciousness (DOC) (Fig. 1, Extended Data Table 1 and Methods). Directly after each olfactory testing session, the state of consciousness of the patient was assessed using standard clinical measures[19,20] to determine whether the patient had unresponsive wakefulness syndrome (UWS) (which has also been referred to as a vegetative state)—in which the patient showed no signs of consciousness[1]—or was in a minimally conscious state (MCS), in which the patient showed inconsistent but reproducible evidence of consciousness[1]. A total of 146 sessions were conducted (1–12 sessions per patient; mean ± s.d. = 3.4 ± 3 sessions; Extended Data Table 1) with inter-session intervals of between 1 and 10 weeks (mean ± s.d. = 2.65 ± 1.7), enabling longitudinal comparisons for 31 patients. Overall, 73 sessions were conducted for 31 patients in a MCS and 73 sessions for 24 patients with UWS (16 of the 24 patients transitioned from UWS to MCS during the study).

We first focused on sensory-driven sniff responses. We compared the normalized nasal inhalation volume (Fig. 1b and Methods) of the first three nasal inhalations after the presentation of unpleasant and pleasant odorants compared with the respiratory baseline during MCS and UWS sessions. Given the abnormal distribution of the data for both MCS and UWS (Shapiro–Wilk tests; first sniff, $W > 0.88$ for all tests, $P < 0.001$ for all tests; second sniff, $W > 0.66$ for all tests, $P < 0.01$ for all tests; third sniff, $W > 0.63$ for all tests, $P < 0.04$ for all tests) and the greater variance in MCS than in UWS (Leven's tests; first sniff, $F_{1,147} = 3.9$, $P = 0.05$; second sniff, $F_{1,147} = 3.3$, $P = 0.07$; third sniff, $F_{1,147} = 1.6$, $P = 0.21$), we used nonparametric tests with Bonferroni correction for multiple comparisons. At the group level, we observed that whereas the nasal

[1]Department of Psychology, University of Cambridge, Cambridge, UK. [2]Azrieli Center for Human Brain Imaging and Research, Weizmann Institute of Science, Rehovot, Israel. [3]Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel. [4]Loewenstein Hospital Rehabilitation Center, Raanana, Israel. [5]Sackler Medical Faculty, Tel-Aviv University, Tel Aviv, Israel. [6]These authors jointly supervised this work: Yaron Sacher, Noam Sobel. ✉e-mail: anat.arzi@gmail.com; noam.sobel@weizmann.ac.il

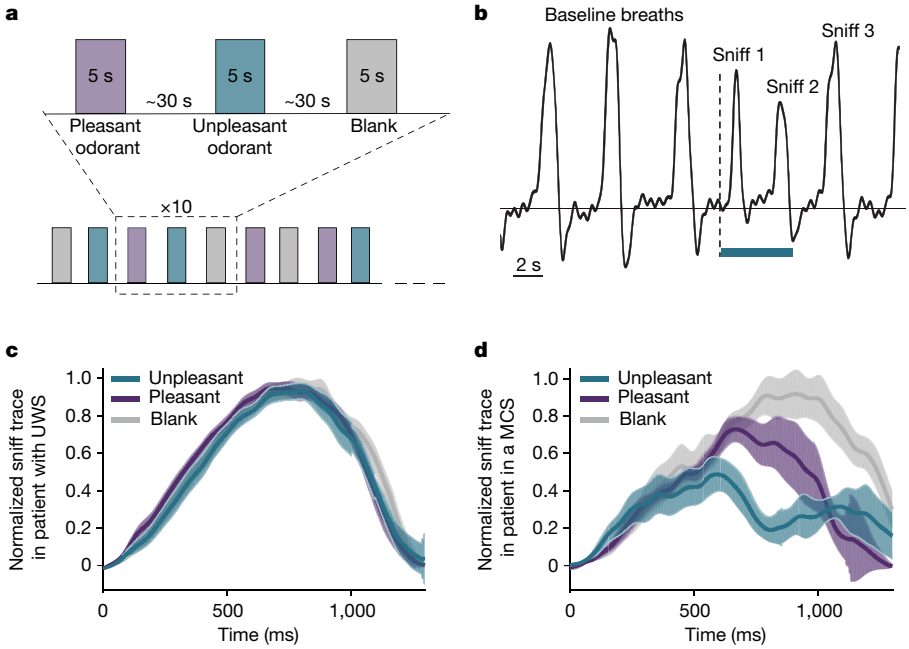


Fig. 1 | Measuring sniff responses in patients with DOC. a, Experimental design for the olfactory test. Pleasant odorants (purple), unpleasant odorants (blue) and blank (grey) were presented in a random order, around 10 times each, for a duration of approximately 5 s. **b**, A trace of nasal respiration during a single trial, during which three baseline respirations and three sniffs after the presentation of the unpleasant odorant were recorded using a nasal cannula connected to a spirometer and amplifier. The dashed line denotes the odorant onset and the blue bar represents the duration of the unpleasant odorant. This trace represents data from patient 37 and is not an illustration. **c**, Example traces lacking sniff responses from patient 28 with UWS. Normalized sniff volumes (mean ± s.d.): pleasant, 1.00 ± 0.08 (11 repetitions); unpleasant, 0.94 ± 0.1 (12 repetitions); blank, 1.00 ± 0.09 (11 repetitions). **d**, Example traces of intact sniff responses from patient 37 in a MCS. Normalized sniff volumes: pleasant, 0.68 ± 0.22 (7 repetitions); unpleasant, 0.56 ± 0.19 (7 repetitions); blank, 0.84 ± 0.21 (6 repetitions). Traces are shown as mean (line) and s.e.m. (shaded areas).

inhalation volume was significantly reduced in response to odorants in MCS sessions, it was not influenced by odorants in UWS sessions. More specifically, at the first sniff after odorant presentation, the normalized sniff volume during MCS sessions decreased from the baseline (which was set to 1) to 0.89 ± 0.18 normalized flow units (NFU) (mean ± s.d.) in response to a pleasant odorant (median = 0.92 NFU, $Z = 5.2$, $P < 0.0001$, 95% confidence interval = 0.85–0.93, effect-size $r = 0.61$; Fig. 2a) and to 0.88 ± 0.2 NFU in response to an unpleasant odorant (median = 0.94 NFU, $Z = 4.7$, $P < 0.0001$, 95% confidence interval = 0.83–0.93, effect-size $r = 0.54$; Fig. 2b). This reflects a reduction of about 10% in nasal airflow that accounts for odorant content in MCS sessions. We did not, however, observe a significant difference between pleasant and unpleasant odorants ($Z = 0.13$, $P = 0.90$, effect-size $r = 0.01$; Fig. 2a, b). In other words, patients showed level-1 (odorant detection) but not level-2 (odorant discrimination) sensory-driven sniff responses during MCS sessions. By contrast, we observed no group-level responses during UWS sessions (normalized nasal inhalation volume, pleasant, mean ± s.d. = 0.97 ± 0.13 NFU, 95% confidence interval = 0.94–0.999, median = 0.99 NFU; unpleasant, mean ± s.d. = 0.97 ± 0.12 NFU, 95% confidence interval = 0.85–0.997, median = 0.98 NFU, difference from baseline, $Z < 2.3$ for all tests, $P > 0.05$ for all Bonferroni-corrected tests, effect-size $r < 0.27$ for all tests; Fig. 2a, b). Furthermore, the responses during UWS sessions were indeed significantly different from those during MCS sessions (difference between groups for the pleasant odorant, $Z = 3.3$, $P = 0.0009$, effect-size $r = 0.39$; and unpleasant odorant, $Z = 2.8$, $P = 0.005$, effect-size $r = 0.33$; Fig. 2a, b). In other words, the level-1 sensory-driven sniff responses that were evident during MCS sessions significantly differentiated these sessions as a group from UWS sessions. Similar results were found for the second sniff after odorant presentation, but not for the third sniff, indicating a genuine, transient, odorant-driven response and not a change in state (Fig. 2e, f and Extended Data Fig. 1). We replicated this analysis, comparing the odorant-driven sniff response to the first inhalation after the presentation of the blank rather than to the respiratory baseline, and obtained similar results (Extended Data Fig. 1 and Supplementary Information). In three controls, we verified that these effects were not a reflection of odorant trigeminality (that is, stinging odours that activate the trigeminal nerve) (Extended Data Fig. 2 and Supplementary Information), tracheostomy of the patient (Extended Data Fig. 3 and Supplementary Information) or the pattern of brain injury (Extended Data Table 2). Together, these results indicate that sensory sniff responses occur in patients in a MCS but not in patients with UWS.

Because consciousness fluctuates in DOC[2], it is common to analyse the data per session and not per patient[21–23]. A case in point is patient 4, who started the study with a session in MCS, then deteriorated, conducting his subsequent session in UWS, only to later recover and conduct a third and final session in MCS (at the time of publication, patient 4 walks and talks). If we were to average the three sessions of patient 4, this would obscure any differences between MCS and UWS. In turn, treating these sessions independently risks overlooking the possible interdependence of measurements obtained from the same individual (albeit in different states). To address interdependence, we reanalysed the data considering only the strongest sniff-response session from each participant who remained unchanged throughout the study. This analysis retained 19 patients in a MCS and 8 patients with UWS. We again observed a significant odorant-induced sniff response in patients in a MCS (sniff 1, pleasant, mean ± s.d. = 0.831 ± 0.18 NFU, 95% confidence interval = 0.75–0.91, median = 0.89 NFU; unpleasant, mean ± s.d. = 0.829 ± 0.21 NFU, 95% confidence interval = 0.73–0.92, median = 0.93 NFU; difference from baseline, $Z > 2.9$ for all tests, $P < 0.004$ for all Bonferroni-corrected tests, effect-size $r > 0.66$ for all tests), yet no sniff response was observed in patients with UWS (sniff 1, pleasant, mean ± s.d. = 0.97 ± 0.11 NFU, 95% confidence interval = 0.90–1.04, median = 0.92 NFU; unpleasant, mean ± s.d. = 0.94 ± 0.06 NFU,

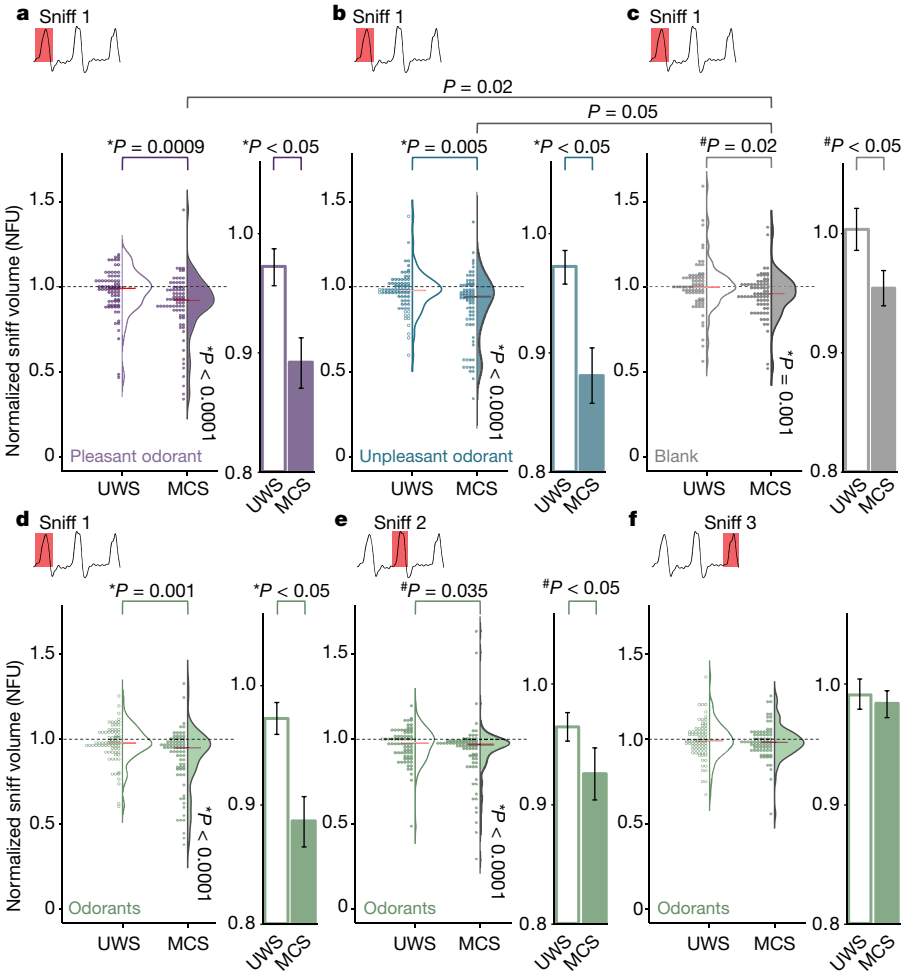


Fig. 2 | The sniff response reflects the level of consciousness in patients with DOC. a–**c**, Normalized sniff volume after a pleasant odorant (**a**; purple), unpleasant odorant (**b**; blue) or the blank (**c**; grey), during UWS sessions (white fill, left trace and graph; $n = 73$) and MCS sessions (colour fill, right trace and graph; $n = 73$) of the first sniff after stimulus delivery. **d**, Combined data from **a** and **b**. **e**, **f**, Combined data from **a** and **b** for the second (**e**) and third (**f**) sniff after odorant delivery (see Extended Data Fig. 1 for separate graphs of pleasant and unpleasant odorants). Left, each dot represents a session, the flat violin plots show the distribution, the red lines denote the median and the dashed horizontal lines denote the baseline value of 1 NFU. Right, data are mean ± s.e.m for each distribution. The *P* values beneath the distributions denote the difference from a baseline inhalation—that is, the existence of a sniff response. The *P* values above the distributions denote the difference in sniff response across groups. *P* values were calculated using two-tailed Wilcoxon signed-rank tests for within-group comparisons and Wilcoxon rank-sum tests for between-group comparisons corrected for multiple comparisons. Corrected *P* values are indicated by an asterisk (*) and uncorrected *P* values are indicated by a hash (#) symbol (see Methods).

95% confidence interval = 0.90–0.98, median = 0.96 NFU, difference from baseline, $P > 0.05$ for all Bonferroni-corrected tests). Moreover, patients in a MCS were not only significantly different from baseline, but also significantly different from the average UWS value (sniff 1 across odorants, $Z = 2.2$, $P = 0.03$, effect-size $r = 0.50$). Thus, sensory sniff responses were evident not only during MCS compared with UWS sessions, but also in patients in a MCS compared with patients with UWS.

We next focused on cognitively driven sniff responses. At the group level, we observed that whereas nasal inhalation volume was significantly reduced in response to blank presentation in MCS sessions, it was uninfluenced by blank presentation in UWS sessions. More specifically, during the first sniff after blank presentation, the normalized sniff volume in MCS sessions decreased from the baseline (which was set to 1) to 0.955 ± 0.13 NFU (mean ± s.d.; 95% confidence interval = 0.93–0.99, median = 0.96 NFU, $Z = 3.25$, $P = 0.001$, effect-size $r = 0.38$; Fig. 2c). This reflects a reduction of about 5% in nasal airflow. By contrast, we observed no response in UWS sessions (blank, mean ± s.d. = 1.00 ± 0.15 NFU, 95% confidence interval = 0.97–1.03, median = 1.00 NFU, $Z = 0.08$, $P = 0.94$, effect-size $r = 0.009$; Fig. 2c). Furthermore, the responses during UWS sessions were significantly different from the MCS sessions (difference between groups, $Z = 2.31$, $P = 0.02$, effect-size $r = 0.27$; Fig. 2c). In other words, the cognitively driven sniff responses that were evident during MCS sessions significantly differentiated these sessions as a group from UWS sessions. This outcome was not evident for the second and third sniffs (Extended Data Fig. 1 and Supplementary Information), again indicating that this difference between groups was a genuine, transient, task-driven response and not a change in state. In summary, we conclude that similar to the level-1 sensory-driven component of the sniff response, the cognitively driven component of the sniff response also reflects the state of consciousness of patients with DOC at the session level of the group.

For this analysis to have value not only for basic science but also for clinical use, it would need to be informative at the level of an individual patient. To make single-patient rather than group judgements, we applied a sniff-response threshold that reflects the extent of change in nasal airflow that constitutes a sniff response within an individual. On the basis of previous studies in healthy participants[17] (Supplementary Table 1), we set a threshold for both the sensory-driven level-1 sniff response (odorant detection) and the cognitive sniff response to a change of more than 15% in the normalized sniff volume between the

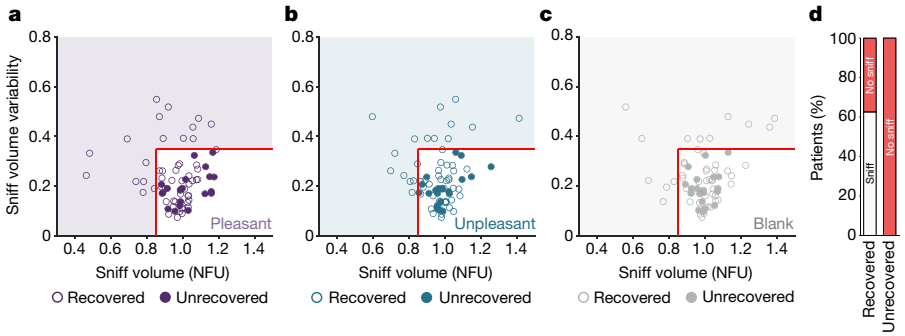


Fig. 3 | The sniff response is associated with the recovery of consciousness in patients with DOC. The red lines denote a previously published sniff-response threshold[11] (more than 15% change in magnitude and/or 0.35 s.d.). Dots within the boxed area (white background; bottom right) reflect sessions without a sniff response; dots outside the boxed area (shaded background) reflect sessions with a sniff response. **a**–**c**, Each dot is a UWS session ($n = 73$). Unfilled dots represent sessions of patients who recovered later (that is, those who transitioned to MCS) and filled dots represent sessions of patients who did not recover consciousness during the study. **a**, Pleasant odorant. **b**, Unpleasant odorant. **c**, Blank. **d**, Percentage of patients with UWS who later transitioned to MCS (left, 'Recovered') or remain unconscious (right, 'Unrecovered') with sniff responses (white; recovered, 62.5%; unrecovered, 0%) and without sniff responses (red; recovered, 37.5%; unrecovered, 100%) across all three conditions.

event and baseline respiration and/or a modulation in sniff volume that reflects a shift in the s.d. of more than 0.35 across trials. The threshold for a sensory-driven level-2 sniff response (odorant discrimination) was set to a change of more than 20% in the normalized sniff volume between pleasant and unpleasant odorants. Using these criteria, we observed that 20 out of 31 patients in a MCS had sniff responses during at least 1 session (19 patients with level-1 sensory-driven sniff responses, of whom 13 had cognitively driven sniff responses, 4 had level-2 sensory-driven sniff responses and 1 had only a strong trend to level-1 sensory-driven sniff responses yet a significant cognitively driven sniff response). This indicates that this measure has a sensitivity of 64.5% to determine a MCS (Extended Data Fig. 4). The false-negative rate reflects the fact that 35.5% of patients in a MCS had no sniff response, indicating that a lack of sniff responses does not necessarily indicate unconsciousness.

We next analysed the patients with UWS. We observed that 9 out of 24 patients with UWS nevertheless had a sensory-driven level-1 sniff response in at least 1 session. Of these, one patient also had a level-2 sniff response. Cognitively driven sniff responses were observed in nine patients with UWS, eight of whom had a sensory-driven sniff response (the ninth patient showed only a trend). Taken together, whereas we failed to observe a sniff response in UWS sessions at the group level, 10 out of 24 individual patients with UWS had sniff responses in at least 1 session. To investigate the implications of the sniff responses that we observed in the patients with UWS, we compared these results to the subsequent clinical progression of the patient over time. Remarkably, we observed that all 10 patients with UWS who had a sniff response in 1 session or more later transitioned to MCS. Thus, a sniff response in UWS indicated transition to MCS with 100% specificity and 62.5% sensitivity (10 out of 16 patients with UWS who transitioned) (Fig. 3a–d and Extended Data Fig. 5). Furthermore, in 4 of the patients the sniff response preceded any other sign of consciousness recovery by days to months (2.5 months, around 2 months, about 1.5 months, and 2 days). These results were not just dependent on the sniff threshold that we used (Extended Data Figs. 6, 7), were not dependent on the odorant condition (Extended Data Fig. 8) and did not reflect the duration of participation (Supplementary Information). Together, this suggests that sniff responses are informative for the prognosis of a patient at the single-patient level.

Lastly, we tested whether sniff responses are informative of the long-term outcome for DOC. We conducted a follow-up investigation at a common point in time about 5 years after injury of the first participant (mean ± s.d. across patients = 37.4 ± 14.7 (range 17–64) months after injury). We observed that whereas patients who had a sniff response shortly after injury generally survived for years, patients who did not have a sniff response after injury generally did not survive during this period. More specifically, only 2 out of 24 patients (8.3%) who had a sniff response after their injury did not survive (these patients died 5 and 7 months after the injury). The remaining 22 out of 24 patients (91.7%) with a sniff response after injury survived (current mean ± s.d. = 37.3 ± 14.1 months after injury). By contrast, 12 out of 19 patients (63.2%) who did not have a sniff response after injury did not survive (patients died within 17.5 ± 12.2 months after injury, median = 13 months; Fig. 4 a–d and Extended Data Fig. 5). Thus, the sensitivity of the sniff response in predicting survival at 37.3 ± 14.1 months after brain injury is 91.7% ($\chi^2 = 14.5$, $P = 0.0001$, effect-size Cramer's $V = 0.45$). Given the strong association between sniff threshold and survival, we would like to reiterate that the sniff-threshold values were set before we obtained the survival rate follow-up data (Supplementary Table 1) and, indeed, this threshold was not the best threshold that could be applied (Extended Data Figs. 6, 7). In addition, we also assessed the functional independence of the 29 surviving patients using the functional independence measure[24]. This measure was independently obtained during clinical assessments of these patients that were conducted at 20.4 ± 11.5 months (mean ± s.d.) after injury. We observed a significant correlation in which the extent of the sensory—but not cognitive—sniff response predicted the later level of independence in patients with UWS (pleasant, Spearman's $r_{39} = -0.49$, $P = 0.001$; unpleasant, Spearman's $r_{39} = -0.60$, $P < 0.0001$; blank, Spearman's $r_{39} = -0.20$, $P = 0.21$ (note that the r values are negative, because a larger sniff response is reflected in a lower post-stimulus volume); Fig. 4e–g and Extended Data Fig. 5). Furthermore, we noticed that this effect was carried by 11 out of 12 surviving patients with UWS who later transitioned to MCS. However, in patients in a MCS, no association was observed (pleasant, Spearman's $r_{62} = 0.30$; unpleasant, Spearman's $r_{62} = 0.24$; blank, $r_{62} = 0.25$, $P > 0.05$ for all Bonferroni-corrected tests). This dissociation between MCS and UWS (difference between correlations, Fisher's exact test, pleasant, $Z = 4.19$, $P < 0.001$; unpleasant, $Z = 4.64$, $P < 0.001$; blank, $Z = 2.27$, $P = 0.012$) in the long-term predictive value of the measurements of functional independence suggests that the sniff response may be informative of the basic mechanisms of life and consciousness, but may not be equally informative for functionality beyond this basic level. Together, these findings suggest that the sniff response can be used as an accessible diagnostic and prognostic tool for the level of consciousness and survival in patients with DOC at the single-patient level.

Olfaction is a primal sensory process in mammals[25] that directly targets the limbic brain without a relay through the thalamus[6]. In humans, the neuroanatomy of the olfactory system is combined with

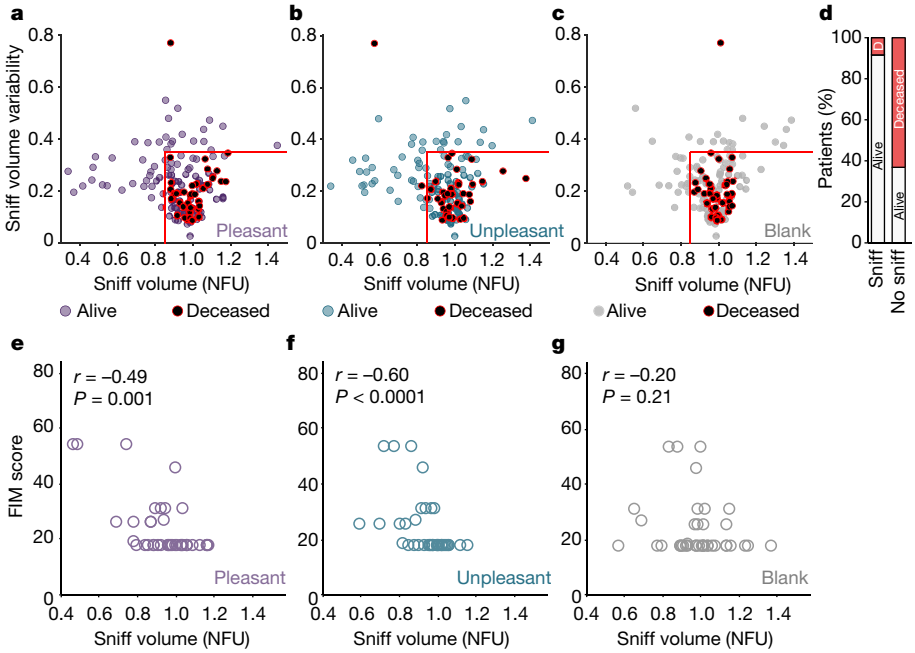


Fig. 4 | The sniff response is associated with long-term survival and functional recovery in patients with DOC. The red lines denote the sniff-response threshold (more than 15% change in magnitude and/or 0.35 s.d.). Dots within the boxed area (bottom right) reflect sessions without a sniff response; dots outside the boxed area lines reflect sessions with a sniff response. **a**–**c**, Each dot is a DOC session (both MCS and UWS; $n = 146$); filled black dots represent sessions of patients who die during the study (deceased) and coloured dots represent sessions of surviving patients (alive; survival (mean ± s.d.), 37.3 ± 14.1 months after brain injury). **a**, Pleasant odorant. **b**, Unpleasant odorant. **c**, Blank. **d**, Percentage of patients with DOC with sniff responses who survived (white, 91.7%) or are deceased (D; red, 8.3%) and patients with DOC without sniff responses who survived (white, 36.8%) or are deceased (red, 63.2%). **e**–**g**, Relation between the functional independence measure (FIM) and normalized sniff volume. Each dot is a UWS session in a surviving patient ($n = 41$). **e**, Pleasant odorant. **f**, Unpleasant odorant. **g**, Blank. Spearman's correlation coefficients (r) are indicated. P values were calculated using Spearman correlations.

aspects of olfactory phenomenology and neurodynamics to afford olfaction a unique position in consciousness[7,26–28]. Olfactory information is processed at several levels and had we relied on higher-order olfactory processing, we would probably not have seen a response in patients with DOC[29]. The olfactory sniff response, however, is a very basic olfactory mechanism that is persistent across species[9,16]. In healthy humans, the olfactory sniff response can persist even without conscious awareness in both wake[14] and sleep (it is strongest during rapid eye movement sleep)[12,18,30]. Nevertheless, such sniff responses rely on intact olfactory neuroanatomy. In this study, we do not claim that this neuroanatomy is the location of consciousness itself, but rather that it probably indicates a level of corticothalamic integrity that is important for consciousness[31,32] and, indeed, for life itself. There are precedents for powerful biomarkers of consciousness in patients with DOC[33], yet the sniff response is also predictive of survival three or more years after injury. The first and last thing that we do in life is to inhale and such inhalations can be modulated by smell, and this modulation is connected with the most basic processes of life.

In addition to the neurological mechanisms of consciousness, our results have clinical implications. Improvements in emergency medicine have increased the survival rates after brain injury but, paradoxically, have resulted in increasing numbers of survivors living with DOC[34,35]. As previously reported[4,5], the rate of misdiagnosis in these patients could be as high as 40%. The sniff response can provide an accessible and easy-to-use tool that may considerably improve this outcome. The ease of use of this tool is an important feature, as it separates this approach from neuroimaging[36,37] and electrophysiology[38,39], which—although powerful for the assessment of consciousness—are not always available to patients with DOC, particularly in developing countries.

The sniff response had a specificity of 100% for the recovery of consciousness. All patients with a sniff response ultimately showed signs of consciousness and all patients with UWS who remained unconscious did not have a sniff response. This places the sniff response among the most-suitable analyses for the estimation of the recovery of consciousness in patients with DOC[33,36–40]. In turn, the sensitivity of the sniff response to detect consciousness in patients in a MCS was 64.5%, and for the detection of the transition from UWS to MCS was 62.5%. These sensitivity rates are higher than most active command-following tests and are similar to other passive paradigms and resting-state studies[33,36–40]. This sensitivity rate, however, suggests that around 35% of conscious patients with DOC had no sniff response. The absence of a sniff response in conscious patients with DOC may reflect a chronic or transient impairment of the olfactory system that is possibly related to their brain injury[41]. Indeed, analysis of the structural brain-imaging data revealed that seven of the eight patients who had at least one session in MCS, but did not have a sniff response, had damage to olfaction-related brain structures (Extended Data Table 2). The absence of a sniff response in conscious patients with DOC may also reflect an inability to execute the precisely timed motor act of sniffing despite intact olfaction[42]. On a related note, non-olfactory volitional nasal inhalations have been used to assess the conscious brain of paralysed individuals in the context of providing a means of communication[43] or as a method for device control[44]. Moreover, non-olfactory volitional nasal inhalation was also tested as a potential method for the detection of consciousness in individuals with DOC; however, most patients with DOC did not inhale on command[45]. Nevertheless, one patient in a MCS in a previous study[45] was able to inhale on command, but could not initiate any other motor movement, further indicating the potential use of this motor response in some cases for which other verifications of response are unavailable. Taken together, we conclude that patients with DOC may not inhale on command[45], but can sniff in response to the presentation of an odorant. This dissociation between volitional

and odorant-induced sniffing further indicates the fundamental role of the olfactory brain in basic mechanisms of arousal.

Finally, in addition to a couple of technical limitations, which are described in the Supplementary Information, we acknowledge a conceptual limitation that is common to studies such as ours. We claim that sniff responses in UWS predicted the later recovery of consciousness. An alternative is that these patients were already conscious, but had been misdiagnosed by the existing standard clinical assessments. The only way to unequivocally settle between these alternatives is by self-reporting of the patients who later transitioned into consciousness. If these patients provide recollections from when they were deemed unconscious, this implies misdiagnosis. Indeed, such instances reveal that in some cases in which all available methods, including functional magnetic resonance imaging and electroencephalograms, determined a lack of conscious awareness, patients nevertheless later recalled events that occurred during their erroneously assessed unconscious state[46,47]. Thus, this fundamental question remains unanswered and is perhaps impossible to answer with the currently available methods. We should emphasize, however, that this does not detract from the contribution of our findings. Whether a sniff response in patients with UWS reflects a better diagnosis or unique prognosis, in both cases it constitutes the signalling of consciousness with all of the associated medical and ethical implications. We conclude that the olfactory sniff response reflects the state of consciousness and is associated with recovery and long-term survival of patients with DOC.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2245-5.

1. Giacino, J. T. et al. Comprehensive systematic review update summary: Disorders of consciousness: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology; the American Congress of Rehabilitation Medicine; and the National Institute on Disability, Independent Living, and Rehabilitation Research. *Neurology* **91**, 461–470 (2018).
2. Giacino, J. T., Fins, J. J., Laureys, S. & Schiff, N. D. Disorders of consciousness after acquired brain injury: the state of the science. *Nat. Rev. Neurol.* **10**, 99–114 (2014).
3. Thibaut, A., Schiff, N., Giacino, J., Laureys, S. & Gosseries, O. Therapeutic interventions in patients with prolonged disorders of consciousness. *Lancet Neurol.* **18**, 600–614 (2019).
4. Stender, J. et al. Diagnostic precision of PET imaging and functional MRI in disorders of consciousness: a clinical validation study. *Lancet* **384**, 514–522 (2014).
5. Schnakers, C. et al. Diagnostic accuracy of the vegetative and minimally conscious state: clinical consensus versus standardized neurobehavioral assessment. *BMC Neurol.* **9**, 35 (2009).
6. Price, J. L. in *The Human Nervous System* (ed. Paxinos, G) (Academic, 1990).
7. Merrick, C., Godwin, C. A., Geisler, M. W. & Morsella, E. The olfactory system as the gateway to the neural correlates of consciousness. *Front. Psychol.* **4**, 1011 (2014).
8. Mainland, J. & Sobel, N. The sniff is part of the olfactory percept. *Chem. Senses* **31**, 181–196 (2006).
9. Wachowiak, M. All in a sniff: olfaction as a model for active sensing. *Neuron* **71**, 962–973 (2011).
10. Frank, R. A., Dulay, M. F. & Gesteland, R. C. Assessment of the sniff magnitude test as a clinical test of olfactory function. *Physiol. Behav.* **78**, 195–204 (2003).
11. Johnson, B. N., Mainland, J. D. & Sobel, N. Rapid olfactory processing implicates subcortical control of an olfactomotor system. *J. Neurophysiol.* **90**, 1084–1094 (2003).
12. Arzi, A. et al. Humans can learn new information during sleep. *Nat. Neurosci.* **15**, 1460–1465 (2012).
13. Rozenkrantz, L. et al. A mechanistic link between olfaction and autism spectrum disorder. *Curr. Biol.* **25**, 1904–1910 (2015).
14. Arzi, A., Rozenkrantz, L., Holtzman, Y., Secundo, L. & Sobel, N. Sniffing patterns uncover implicit memory for undetected odors. *Curr. Biol.* **24**, R263–R264 (2014).
15. Reden, J., Draf, C., Frank, R. A. & Hummel, T. Comparison of clinical tests of olfactory function. *Eur. Arch. Otorhinolaryngol.* **273**, 927–931 (2016).
16. Rojas-Líbano, D. & Kay, L. M. Interplay between sniffing and odorant sorptive properties in the rat. *J. Neurosci.* **32**, 15577–15589 (2012).
17. Johnson, B. N., Russell, C., Khan, R. M. & Sobel, N. A comparison of methods for sniff measurement concurrent with olfactory tasks in humans. *Chem. Senses* **31**, 795–806 (2006).
18. Arzi, A. et al. Olfactory aversive conditioning during sleep reduces cigarette-smoking behavior. *J. Neurosci.* **34**, 15382–15393 (2014).
19. Giacino, J. T. & Kalmar, K. *Coma Recovery Scale-Revised: Administration and Scoring Guidelines*. https://www.tbims.org/combi/crs/CRS%20Syllabus.pdf (JFK Johnson Rehabilitation Institute and New Jersey Neuroscience Institute, JFK Medical Center, 2004).
20. Rappaport, M., Dougherty, A. M. & Kelting, D. L. Evaluation of coma and vegetative states. *Arch. Phys. Med. Rehabil.* **73**, 628–634 (1992).
21. Bareham, C. A. et al. Longitudinal bedside assessments of brain networks in disorders of consciousness: case reports from the field. *Front. Neurol.* **9**, 676 (2018).
22. Wannez, S., Heine, L., Thonnard, M., Gosseries, O. & Laureys, S. The repetition of behavioral assessments in diagnosis of disorders of consciousness. *Ann. Neurol.* **81**, 883–889 (2017).
23. Bareham, C. A. et al. Longitudinal assessments highlight long-term behavioural recovery in disorders of consciousness. *Brain Commun.* **1**, fcz017 (2019).
24. Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V. & Hamilton, B. B. The structure and stability of the functional independence measure. *Arch. Phys. Med. Rehabil.* **75**, 127–132 (1994).
25. Rowe, T. B., Macrini, T. E. & Luo, Z.-X. Fossil evidence on origin of the mammalian brain. *Science* **332**, 955–957 (2011).
26. Sela, L. & Sobel, N. Human olfaction: a constant state of change-blindness. *Exp. Brain Res.* **205**, 13–29 (2010).
27. Keller, A. Attention and olfactory consciousness. *Front. Psychol.* **2**, 380 (2011).
28. Stevenson, R. J. Phenomenal and access consciousness in olfaction. *Conscious. Cogn.* **18**, 1004–1017 (2009).
29. Nigri, A. et al. Central olfactory processing in patients with disorders of consciousness. *Eur. J. Neurol.* **23**, 605–612 (2016).
30. Arzi, A. et al. The influence of odorants on respiratory patterns in sleep. *Chem. Senses* **35**, 31–40 (2010).
31. Forgacs, P. B. et al. Preservation of electroencephalographic organization in patients with impaired consciousness and imaging-based evidence of command-following. *Ann. Neurol.* **76**, 869–879 (2014).
32. Fernández-Espejo, D. et al. A role for the default mode network in the bases of disorders of consciousness. *Ann. Neurol.* **72**, 335–343 (2012).
33. Noirhomme, Q., Brecheisen, R., Lesenfants, D., Antonopoulos, G. & Laureys, S. "Look at my classifier's result": disentangling unresponsive from (minimally) conscious patients. *Neuroimage* **145**, 288–303 (2017).
34. Majdan, M. et al. Epidemiology of traumatic brain injuries in Europe: a cross-sectional analysis. *Lancet Public Health* **1**, e76–e83 (2016).
35. Aidinoff, E. et al. Vegetative state outcomes improved over the last two decades. *Brain Inj.* **32**, 297–302 (2018).
36. Monti, M. M. et al. Willful modulation of brain activity in disorders of consciousness. *N. Engl. J. Med.* **362**, 579–589 (2010).
37. Coleman, M. R. et al. Towards the routine use of brain imaging to aid the clinical diagnosis of disorders of consciousness. *Brain* **132**, 2541–2552 (2009).
38. Napolitani, M. et al. Transcranial magnetic stimulation combined with high-density EEG in altered states of consciousness. *Brain Inj.* **28**, 1180–1189 (2014).
39. Engemann, D. A. et al. Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain* **141**, 3179–3192 (2018).
40. Bekinschtein, T. A. et al. Classical conditioning in the vegetative and minimally conscious state. *Nat. Neurosci.* **12**, 1343–1349 (2009).
41. Green, P., Rohling, M. L., Iverson, G. L. & Gervais, R. O. Relationships between olfactory discrimination and head injury severity. *Brain Inj.* **17**, 479–496 (2003).
42. Sobel, N. et al. An impairment in sniffing contributes to the olfactory impairment in Parkinson's disease. *Proc. Natl Acad. Sci. USA* **98**, 4154–4159 (2001).
43. Plotkin, A. et al. Sniffing enables communication and environmental control for the severely disabled. *Proc. Natl Acad. Sci. USA* **107**, 14413–14418 (2010).
44. Haviv, L. et al. Using a sniff controller to self-trigger abdominal functional electrical stimulation for assisted coughing following cervical spinal cord lesions. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 1461–1471 (2017).
45. Charland-Verville, V. et al. Detection of response to command using voluntary control of breathing in disorders of consciousness. *Front. Hum. Neurosci.* **8**, 1020 (2014).
46. Owen, A. *Into the Gray Zone: A Neuroscientist Explores the Border Between Life and Death* (Simon and Schuster, 2017).
47. Kondziella, D., Friberg, C. K., Frokjaer, V. G., Fabricius, M. & Møller, K. Preserved consciousness in vegetative and minimal conscious states: systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatry* **87**, 485–492 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Patients

In total, 50 patients (age, 43.4 ± 17 years (mean ± s.d.; 9 women) were recruited over approximately 4 years at the Intensive Care and Rehabilitation of Consciousness Department at the Loewenstein Hospital Rehabilitation Center, Raanana, Israel. Patients were tested during multiple sessions (range, 1–13 sessions; mean ± s.d., 3.8 ± 2.98 sessions; total, 190 sessions) separated by days to weeks depending on their clinical and personal availability. Out of 190 olfactory testing sessions, 41 sessions were excluded due to a lack of stable nasal respiration and 3 sessions were excluded due to a nasal inhalation volume that was larger than 3.5 s.d. of the group mean (see 'Inclusion and exclusion criteria'), retaining 146 sessions from 43 patients (Extended Data Table 1). The study was approved by the ethics committee of Loewenstein Hospital Rehabilitation Center. Written informed consent was obtained from the legal guardian of the patient.

Measurement of nasal airflow

We used methods we have applied extensively in previous studies[11,13,17,43,44]. In brief, during the olfactory testing sessions, the nasal airflow of patients was measured using a nasal cannula (1103, Teleflex Medical) linked directly to a spirometer (ML141, AD Instruments; H_2O resolution, 15.6 μV) that converted airflow into a voltage, which was sent to an instrumentation amplifier (PowerLab 16SP Monitoring System, AD Instruments) sampling at 1,000 Hz using LabChart software (AD Instruments).

Odorants

We used two odorant mixtures (a pleasant 'shampoo' and unpleasant 'rotten fish' scent, both from Sensale) that were presented to all patients, and two pure odorant molecules (the pleasant phenylethyl alcohol (PEA), CAS 102-20-5, which smells like roses; and the slightly unpleasant decanoic acid, CAS 334-48-5, which smells like crayons; both from Sigma-Aldrich) that were presented to a subset of 31 patients (Extended Data Fig. 2). These odorants were effective in generating sniff responses in previous studies[8,12,13,18]. The odorants were absorbed in a cotton pad and placed in a sniff jar. A jar with only a cotton pad served as the blank.

Procedure

At the beginning of each session the experimenter explained to the patient that odorants will be presented using sniff jars, and that nasal respiration will be monitored during the session. This was repeated for each session despite the researcher having no indication whether the patient heard or understood what was said. Next, the experimenter gently applied a nasal cannula to the patient's nostrils to record nasal respiration. If no nasal respiration was observed, the experimenter waited for a few minutes before another observation attempt was made. If there were still no signs of nasal respiration, the experimenter changed the body position of the patient if possible. If no nasal respiration was observed after the change in position, the olfactory testing session was terminated. Lack of nasal respiration in a given session did not exclude the patient from future sessions. If nasal respiration was detected, the olfactory testing session began. During each trial a jar with a pleasant odorant, an unpleasant odorant or the blank was presented to the patient for approximately 5 s. The jar was brought into the field of view of the patient and was then placed under the nose of the patient (without touching the patient) at the exact end of an exhale so that the patient would receive the odorant in the following inhale. Each odorant and the blank jar were presented around 10 times in a random order as long as nasal respiration was evident. In rare cases, due to clinical needs or the lack of stable nasal respiration, the session ended before the administration of all 10 repetitions. Following each olfactory testing session, we obtained a behavioural clinical evaluation of the state of consciousness of the patient to determine whether the session was in UWS or MCS (including emergence from MCS).

Behavioural clinical evaluation

After each olfactory testing session, the state of consciousness of the patient was evaluated using the Coma Recovery Scale Revised (CRS-R)[19] and/or the Coma–Near Coma (CNC) scale[20]. The CRS-R evaluates the presence or absence of responses to auditory, visual, motor, oromotor, communication and arousal functions. CRS-R is both a quantitative assessment—with scores ranging from 0 (lowest level of consciousness) to 23 (highest level of consciousness)—and a qualitative assessment—with 4 levels comprising coma, UWS, MCS and emergence from MCS for which specific behaviours define each level. The CNC evaluates the occurrence of responses to visual, auditory, command following, threat response, olfactory, tactile, pain and vocalization. The CNC is both a quantitative assessment—with scores ranging from 4 (lowest level of consciousness) to 0 (highest level of consciousness)—and a qualitative assessment—with 5 levels comprising extreme coma (3.5–4), marked coma (2.9–3.49), moderate coma (2.01–2.89), near coma (0.9–2), no coma (0–0.89). We converted the CNC qualitative levels, on the basis of a previous study[20], as follows: extreme coma and marked coma, UWS; moderate coma and near coma, MCS; no coma, emergence from MCS. In addition, all patients were periodically (unrelated to our testing schedule) assessed using the Loewenstein Communication Scale (LCS)[48]. The LCS evaluates five hierarchical functions: mobility, respiration, visual responsiveness, auditory comprehension and linguistic skills (verbal or alternative). The LCS is quantitative with scores ranging from 0 to 100, for which scores of up to 20 are considered UWS and scores above 20 are considered MCS. Evaluations of the state of consciousness obtained directly after the olfactory testing session were missing in three cases owing to technical errors and were estimated based on the nearest LCS (all three patients were in a MCS). Follow-up functional independence of the patients was evaluated using the functional independence measure[24]. The functional independence measure is quantitative with scores ranging from 18 to 126.

We evaluated 21 patients using only CNC and LCS, the state of consciousness of the remaining patients was evaluated using CRS-R, CNC and LCS. In the cases in which both CRS-R and CNC were used, the level of consciousness was determined by the CRS-R[49]. To estimate the effect of having only CNC and LCS estimates, we compared the classification of UWS and MCS between CRS-R and CNC scales in the current dataset. We observed disagreement in 30 out of 80 sessions with both scales (in 29 sessions CRS-R indicated UWS and the CNC suggested MCS; in one session the CRS-R indicated MCS and the CNC suggested emergence from MCS). This could suggest that the CNC division to UWS (extreme coma and marked coma) and MCS (moderate coma and near coma) using the scale subcategories might be too liberal. Out of the 21 patients assessed only with CNC scale after the olfactory session, 4 were in UWS during all sessions, 12 were in MCS during all sessions, and 5 transitioned between UWS and MCS across sessions. If indeed the CNC is too liberal, it is possible that some patients in a MCS were in fact UWS. The consequences of such misclassification on the group-level analysis means that we could have underestimated the observed findings. Thus, we conclude that although such misclassification would be unfortunate, it would not weaken, but only strengthen our effects. A second possible consequence of the use of the CNC scale is the detection of recovery while no recovery occurred. Out of the 16 patients who transited from UWS to MCS, 5 were assessed using the CNC but not with the CRS-R. Notably, the LCS[48]—which was assessed independently by the hospital team—provided additional evidence for conscious awareness in all of five of the patients, thus supporting the CNC behavioural assessment. Therefore, the lower sensitivity of CNC versus CRS-R in this study may have underestimated the power of the results, but does not appear to inaccurately detect the transition to MCS.

Airflow analysis

The ongoing respiration trace was filtered as follows. We first applied an equiripple low-pass filter of 10 Hz (pass frequency, 10 Hz; stop

frequency, 20 Hz; allowed ripple amplitude, 1 dB; stop amplitude attenuation, 60 dB; number of coefficients, 224). Next, for the identification of inhales and exhales, the hysteresis was applied after filtering the signal using an equiripple low-pass filter of 5 Hz (pass frequency, 5 Hz; stop frequency, 6 Hz, allowed ripple amplitude, 1 dB; stop amplitude attenuation, 60 dB; number of coefficients, 1,975). Discrete inhales and exhales were identified on the basis of hysteresis of either up to 5 mV or 5% of the difference between the minimum and maximum values, whichever was the smallest of the two, and a minimum duration of 250 ms. If the respiration variability was high, the hysteresis value was not constant for the whole session but was calculated using a sliding window of 30 s, based on the respiration variance. To account for changes in the respiration pattern across a session and between sessions, each nasal inhalation following a stimulus was normalized by dividing the nasal inhalation volume after the stimulus by the baseline inhalation volume (an average of three inhalations before odour administration; Fig. 1b). As patients with DOC often breathe through a tracheostomy tube, we tested whether tracheostomy modulated the results. We found that although tracheostomy significantly reduces nasal inhalation (not normalized), it does not modulate normalized sniff responses (Extended Data Fig. 3).

Definition of the sniff-response threshold

At the single-patient level, the sensory-driven level-1 sniff-response and cognitively driven sniff-response thresholds were defined based on previous studies in healthy participants[17] as: (1) a reduction in sniff-response magnitude of 15% or more in relation to baseline respiration (Supplementary Table 1) and/or (2) a sniff-response s.d. across all trials in the session of more than 0.35, based on the variability within the data-set (twice the averaged s.d. in MCS sessions). The level-2 sniff-response thresholds were defined as (1) a 20% difference in sniff-response magnitude between pleasant and unpleasant odorants and (2) a reduction in the normalized nasal inhalation volume in relation to baseline (<1) for both odorants. More specifically, the sniff-response magnitude threshold was based on previous research investigating methods of sniff measurements in healthy participants[17]. To define the sniff-response threshold, we calculated the change in sniff volume averaged across pleasant (phenylethyl alcohol; CAS 102-20-5, Sigma-Aldrich) and unpleasant (valeric acid; CAS 109-52-4, Sigma-Aldrich) odorants in relation to clean air in healthy participants[17] (Supplementary Table 1). Sniff volume (integral) values used for the calculation were measured using the same method as in our study (nasal cannula) and were the following: sniff volume for pleasant odorants, 0.837; sniff volume for unpleasant odorants, 0.641; nasal inhalation volume for clear air, 0.86. This leads to an approximately 15% decrease if using the formula $[(0.837 + 0.641)/2]/0.86 = 0.8593$.

Inclusion and exclusion criteria

Sessions. A session was excluded from the group analysis on the basis of the following criteria. (1) Nasal respiration was not evident or unstable, and therefore no trials or not enough trials were available (41 sessions). Only sessions with more than 15 trials were included in the analysis (7 patients had no nasal respiration or unstable nasal respiration and were therefore not included in the analysis). (2) The averaged normalized nasal inhalation volume was larger than 3.5 s.d. of the group mean. One session from each of three different patients was excluded using this criterion. Notably, all 3 patients had a later score of at least 48 (48, 62 and 84) in the functional independence measure[24] and at least 49 (49, 55 and 57) in the LCS[48], indicating the emergence from MCS. This suggests that nasal inhalation could be informative of consciousness even in cases of potentially altered olfaction.

These exclusion criteria retained 43 out of 50 patients with DOC (that is, 14% patients were excluded) (Extended Data Table 1), 146 out of 190 sessions (that is, 23.1% were sessions excluded) and 5,934 out of 6,106 trials (that is, 2.82% trials were excluded) in this study.

Trials. A trial was excluded from a session on the basis of the following criteria. (1) Baseline inhalation was unstable, presenting a monotonic decrease or increase of at least 40% change in the peak between the first and third baseline inhalation, and at least a 25% change between the first and second, and between the second and third inhalations. (2) No sniff was detected within 6.5 s of odour presentation. (3) No baseline inhalation was detected—respiration was too flat or if two trials were too close in time and therefore there were only three or less inhalations between trials. (4) There was an extreme change in the inhalation volume between baseline and the subsequent inhalations. To identify these rare cases (12 trials, 0.19% of all trials), we measured the average and s.d. of inhale volumes at baseline and after the stimulus, and the coefficients of variation (the ratio of the s.d. to mean). A trial was excluded if the following three conditions all occurred. (a) The maximum of the two coefficients of variation was below 20%; (b) the percentage signal change of the two coefficients of variation was below 50%; and (c) the percentage signal change of the two means was above 50%.

These exclusion criteria retained 5,778 out of the remaining 5,934 trials after session exclusion (that is, 2.63% excluded) in this study.

Baseline inhalations. A baseline inhalation was excluded from the averaged baseline inhalation on the basis of the following criteria. (1) The baseline inhale started more than 30 s before odour presentation. (2) The baseline inhale overlapped with the sniff response of a previous trial. (3) The inhalation volume was 25% smaller or larger than the other two baseline inhalations in the trial. In this case, the baseline inhalation with the maximal difference in inhale volumes from the median volume was excluded.

These exclusion criteria retained 16,300 out of 17,334 baseline inhalations (that is, 5.96% excluded) in this study.

Sniffs. A sniff was excluded on the basis of the following criteria. (1) The normalized sniff volume was ±3.5 s.d. of the averaged sniff in the session. (2) The first sniff was excluded if no sniff was detected within 6.5 s of odour presentation (and then the trial was excluded as well), and all other sniffs were excluded if not detected within 6.5 s of the previous sniff (and then the later sniffs were also excluded). In 8 patients, the respiration rate was slower than typical and the threshold was extended (7.5 s in 3 patients, 8.5 s in 3 patients and 11 s in 1 patient).

These exclusion criteria retained 16,999 out of 17,334 sniffs (that is, 1.93% excluded) in this study.

Sensitivity and specificity

Specificity (true-negative rate) scores were calculated for patients with UWS who remained UWS. Sensitivity (true-positive rate) scores were calculated for patients with UWS who transitioned to MCS, for the long-term survival rates of a patient and for patients in a MCS[33].

Statistical analysis

Normalized nasal inhalation volume values were not normally distributed (Shapiro–Wilk tests; first sniff, $W > 0.88$ for all tests, $P < 0.001$ for all tests; second sniff, $W > 0.66$ for all tests, $P < 0.01$ for all tests; third sniff, $W > 0.63$ for all tests, $P < 0.04$ for all tests) and displayed greater variance in MCS than in UWS sessions (Leven's tests; first sniff, $F_{1,147} = 3.9$, $P = 0.05$; second sniff, $F_{1,147} = 3.3$, $P = 0.07$; third sniff, $F_{1,147} = 1.6$, $P = 0.21$). Thus, nonparametric tests were used. For statistical analysis between MCS and UWS sessions, Wilcoxon rank-sum tests were used and for statistical analysis within each group, Wilcoxon signed-rank tests were used. Bonferroni corrections for multiple comparisons were applied for comparisons of nasal inhalation volume between MCS and UWS sessions and also within a group (two odours × three sniffs: $0.05/6 = 0.0083$). The effect size of nonparametric, dependent samples was calculated using the formula $r = Z/\sqrt{(n)}$ as previously described[50], where Z is the Wilcoxon signed-rank statistic and n is the sample size.

The effect size for nonparametric, independent samples was estimated using Cliff's δ[51]. The effect size for χ^2 tests was estimated using Cramer's V[52]. The relation between the sniff responses and functional independence measure[24] was assessed using Spearman correlations and was Bonferroni corrected for multiple comparisons (two states × three sniffs: 0.05/6 = 0.0083). Three sessions from two patients who had outlier values in the sniff response in later sessions, suggesting impaired olfaction, were excluded from the correlation analysis. When including these three sessions similar results were obtained (pleasant $r_{42} = -0.27$, $P = 0.075$; unpleasant, $r_{42} = -0.45$, $P = 0.002$; blank, $r_{42} = -0.04$, $P = 0.80$).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Respiration data that support the findings of this study are available from GitLab (https://gitlab.com/liorg/OlfactorySniffingAnalysis/). Source data for Figs. 1–4 are provided with the paper.

Code availability

Custom code created and used in this study is available from GitLab (https://gitlab.com/liorg/OlfactorySniffingAnalysis/).

48. Borer-Alafi, N., Gil, M. Sazbon, L. & Korn, C. Loewenstein communication scale for the minimally responsive patient. *Brain Inj.* **16**, 593–609 (2002).
49. Seel, R. T. et al. Assessment scales for disorders of consciousness: evidence-based recommendations for clinical practice and research. *Arch. Phys. Med. Rehabil.* **91**, 1795–1813 (2010).
50. Rosenthal, R. in *The Handbook of Research Synthesis* (eds Cooper, H. & Hedges, L. V.) 231–244 (Russell Sage Foundation, 1994).
51. Cliff, N. Dominance statistics: ordinal analyses to answer ordinal questions. *Psychol. Bull.* **114**, 494–509 (1993).
52. Kim, H.-Y. Statistical notes for clinical researchers: chi-squared test and Fisher's exact test. *Restor. Dent. Endod.* **42**, 152–155 (2017).

Acknowledgements We thank the patients, their families and their caregivers for their cooperation; and O. Perl, L. Rösler and A. Elite for discussions. Work in the Sobel laboratory is supported by the Rob and Cheryl McEwen Fund for Brain Research. A.A. is supported by the Blavatnik Family Foundation, a Royal Society – Kohn International fellowship (NF150851) and an European Molecular Biology Organization (EMBO) fellowship (ALTF 33-2016).

Author contributions Conceptualization: A.A. and N.S.; data collection: A.A., L.R. and L.G.; analysis: A.A., L.R., L.G., A.R. and N.S.; funding acquisition: A.A., N.S. and Y.S.; investigation: A.A., L.R., D.R. and Y.H.; methodology: A.A., L.R., L.G., A.R. and N.S.; project administration: A.A., L.R., D.R., Y.H., G.C., T.G., B.-Z.K., A.O., E.A., Y.S. and N.S.; resources: A.A., L.R., L.G., D.R., Y.H., G.C., T.G., B.-Z.K., A.O., E.A., Y.S. and N.S.; software: A.A., L.R. and L.G.; supervision: A.A., T.A.B., T.G., B.-Z.K., A.O., E.A., Y.S. and N.S.; validation: A.A., L.R. and L.G.; visualization: A.A. and N.S.; writing of the original draft: A.A. and N.S.; writing, review and editing: A.A., L.R., L.G., D.R., Y.H., T.A.B., G.C., T.G., B.-Z.K., A.O., E.A., Y.S. and N.S.

Competing interests The findings of this manuscript are being used by the Weizmann Institute Office of Technology Licensing (Yeda) for the submission of a patent for a method of the detection of consciousness.

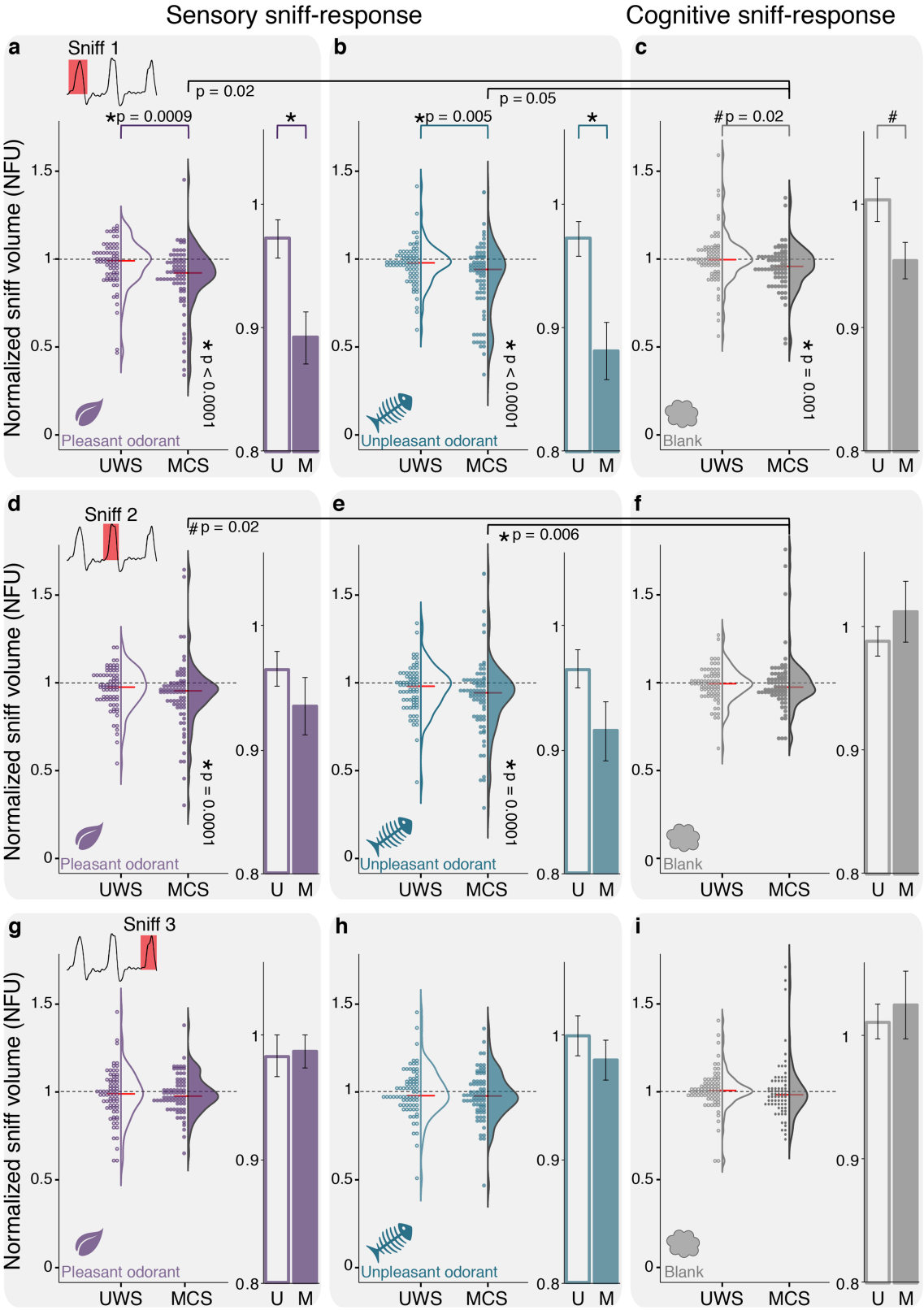
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2245-5.

Correspondence and requests for materials should be addressed to A.A. or N.S.

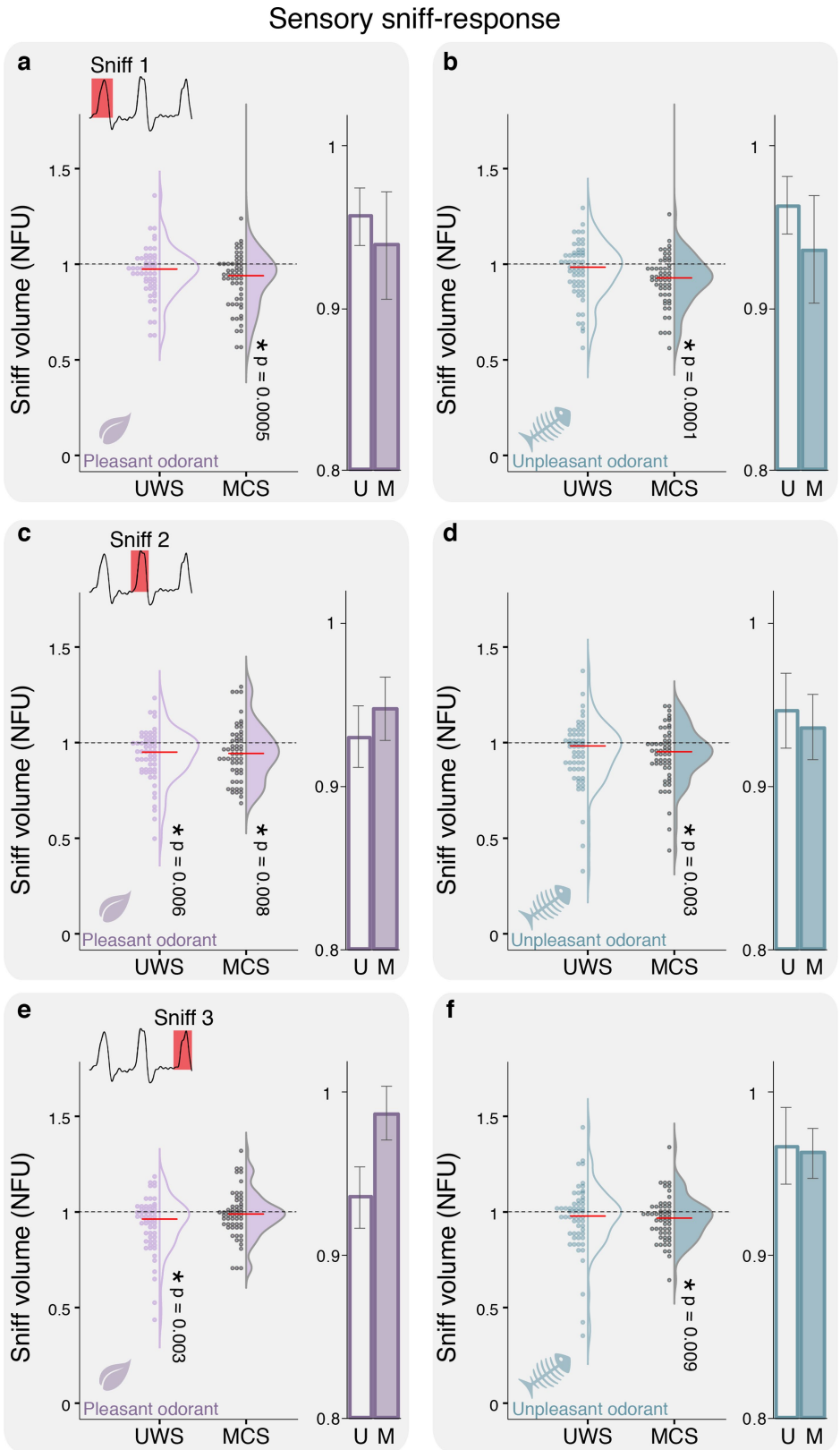
Peer review information *Nature* thanks Hakwan Lau, Johan Lundström, Nicholas Schiff and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

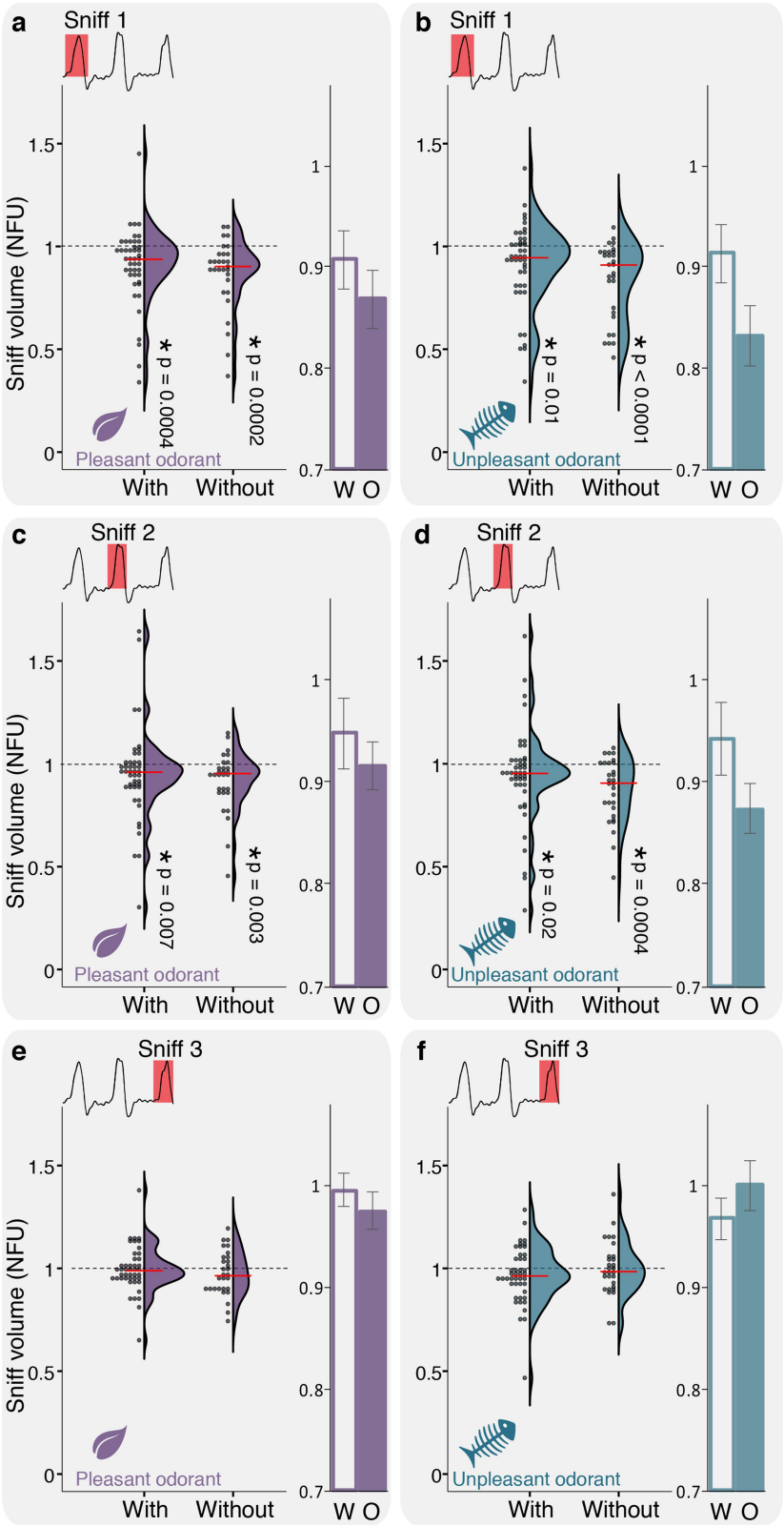


Extended Data Fig. 1 | See next page for caption.

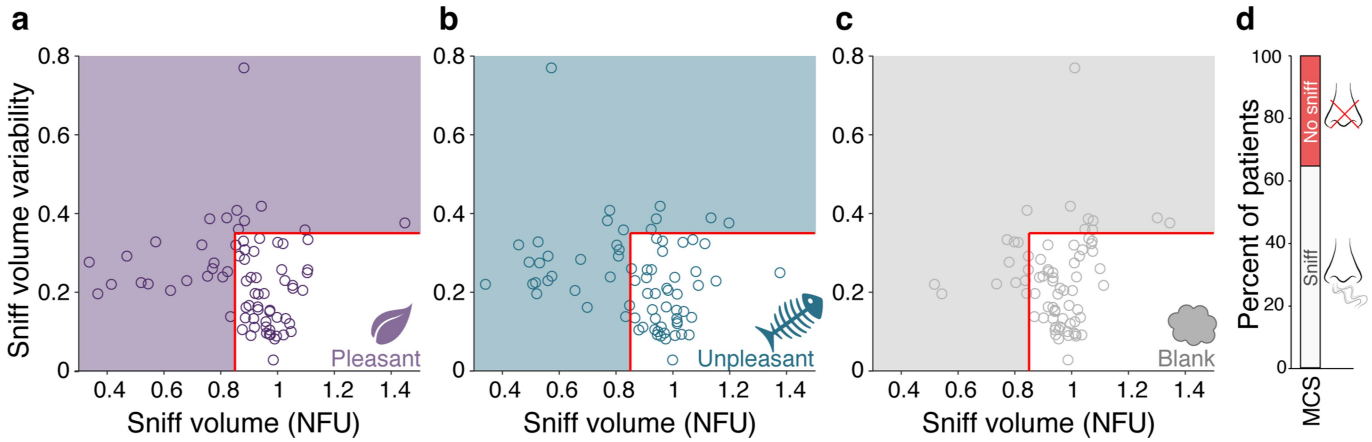
Extended Data Fig. 1 | The sniff response reflects the current level of consciousness in patients with DOC. Data are displayed by odorant and sniff. **a**–**i**, Normalized sniff volume after pleasant odorants (**a**, **d**, **g**), unpleasant odorants (**b**, **e**, **h**) and blank (**c**, **f**, **i**) during UWS (U) sessions (outline; $n = 73$) and MCS (M) sessions (filled; $n = 73$) for the first (**a**–**c**), second (**d**–**f**) and third (**g**–**i**) sniff after stimulus delivery. Left, each dot represents a session; flat violin plots show the distribution; the red lines denote the median; and the dashed horizontal lines denote the baseline value at 1 NFU. Right, data are the mean ± s.e.m. for each distribution. The *P* values beneath the distribution denotes its difference from baseline inhalation. The *P* values above the distributions denote the difference in sniff response across groups. *P* values were calculated using two-tailed Wilcoxon signed-rank tests for within-group comparisons and Wilcoxon rank-sum tests for between-group comparisons corrected for multiple comparisons. Corrected *P* values are indicated by an asterisk (*) and uncorrected *P* values are indicated by a hash (#) symbol (see Methods). $^{*}P < 0.05$; $^{\#}P < 0.05$. Further analyses are provided in the Supplementary Information.



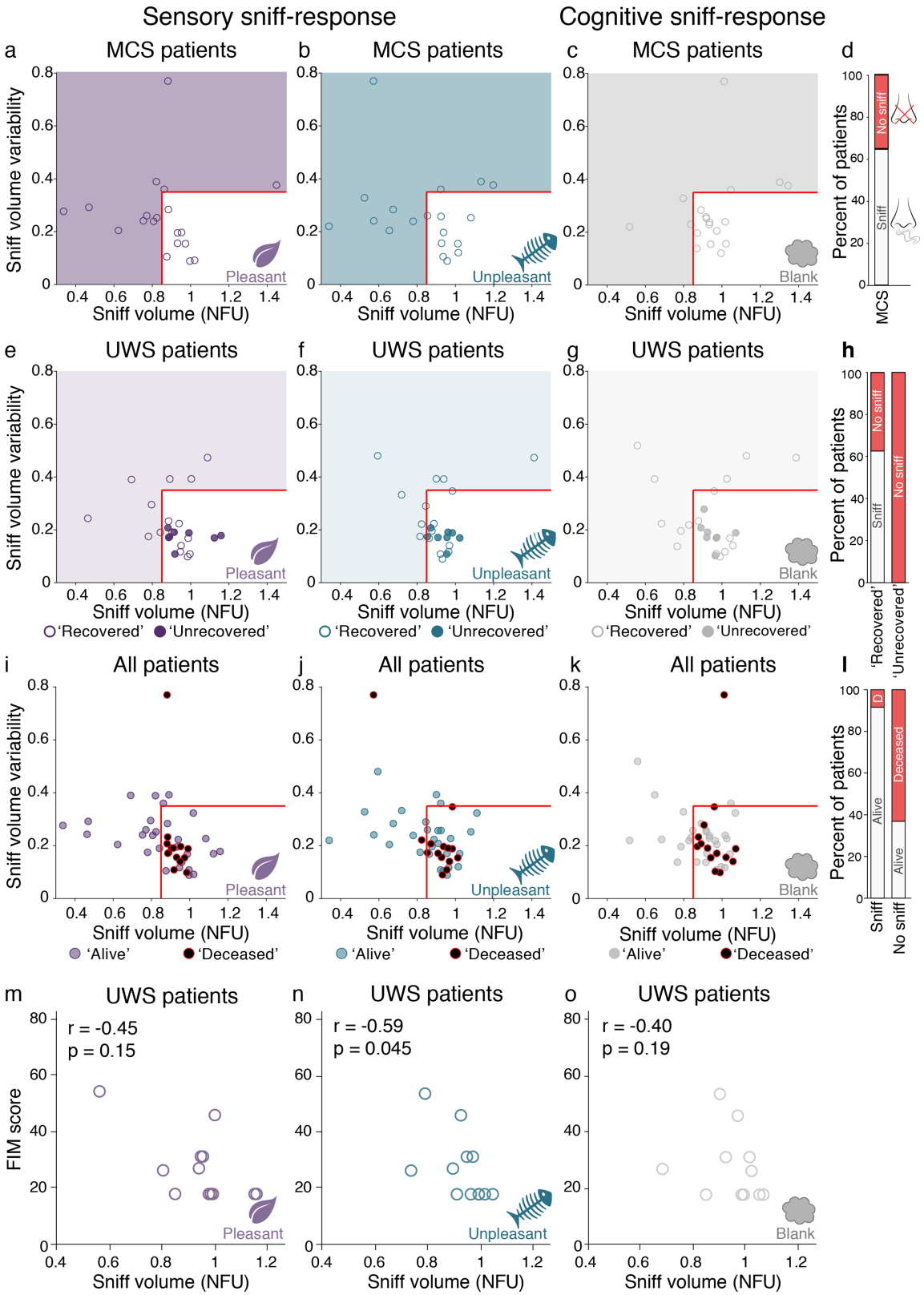
Extended Data Fig. 2 | The sniff response to pure olfactory odorants in patients with DOC. To estimate whether the effects that we observed were dependent on the contribution of the trigeminal nerve, we exposed a subset of patients to pure olfactory odorants and observed the replication of the effects. The normalized sniff volume after exposure to the pure olfactants of the pleasant odorant (phenylethyl alcohol; **a**, **c**, **e**) and unpleasant odorant (decanoic acid; **b**, **d**, **f**) during UWS sessions (outline; pleasant, $n = 56$; unpleasant, $n = 57$) and MCS sessions (filled; pleasant, $n = 56$; unpleasant $n = 56$) for the first (**a**, **b**), second (**c**, **d**) and third (**e**, **f**) sniff after odorant delivery. Left, each dot represents a session; flat violin plots show the distribution; the red lines denote the median; and the dashed horizontal lines denote the baseline value at 1 NFU. Right, data are the mean ± s.e.m. for each distribution. The *P* value beneath the distribution denotes its difference from baseline inhalation—that is, the existence of a sniff response. *$P < 0.05$; two-tailed Wilcoxon test. Further analyses are provided in the Supplementary Information.



Extended Data Fig. 3 | The sniff response is similar with and without tracheostomy. Normalized sniff volume after pleasant (**a**, **c**, **e**) or unpleasant (**b**, **d**, **f**) odorants during MCS sessions with (W; $n = 44$) and without (O; $n = 29$) tracheostomy for the first (**a**, **b**), second (**c**, **d**) and third (**e**, **f**) sniff after odorant delivery. Left, each dot represents a session; flat violin plots show the distribution; the red lines denote the median; and the dashed horizontal lines denote the baseline value at 1 NFU. Right, data are the mean ± s.e.m. for each distribution. The *P* value beneath the distribution denotes its difference from baseline inhalation—that is, the existence of a sniff response. *$P < 0.05$; two-tailed Wilcoxon test. About 60% of MCS sessions and 80% of UWS sessions were conducted in patients with a tracheostomy. Although tracheostomy significantly reduces nasal airflow, a measurable portion of nasal airflow remains. For example, we note that the raw data in Fig. 1c, d were obtained with a tracheostomy. Further analyses are provided in the Supplementary Information.

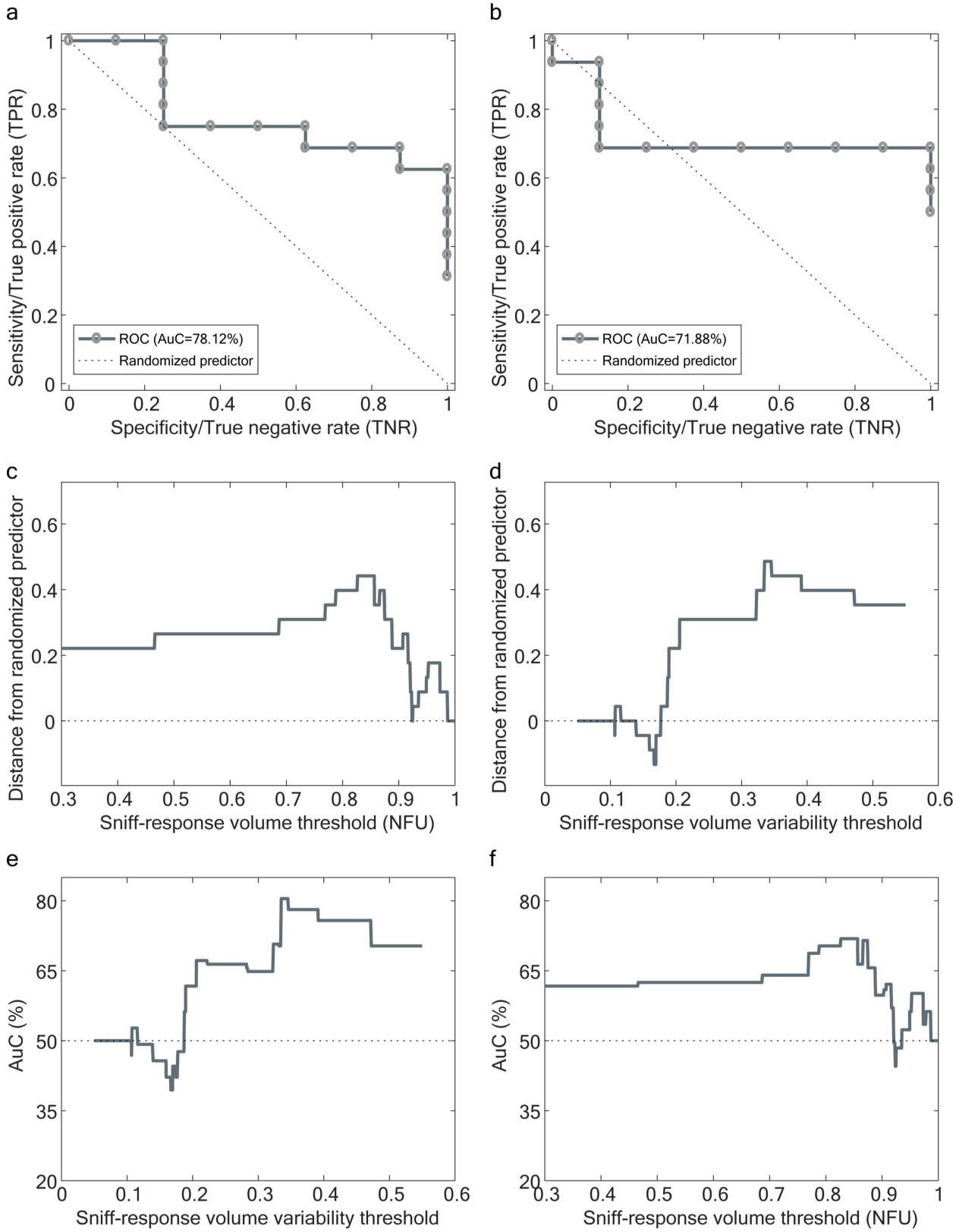


Extended Data Fig. 4 | The sniff response during MCS sessions. The red lines denote the sniff-response threshold (more than 15% change in magnitude and/or 0.35 s.d.): dots within the lines (white background) reflect sessions without a sniff response; dots beyond the lines (shaded background) reflect sessions with a sniff response. **a**–**c**, Each dot is a MCS session. **a**, Pleasant odorant. **b**, Unpleasant odorant. **c**, Blank. **d**, Percentage of patients in a MCS (not sessions) with sniff responses (white, 64.5%) and without sniff responses (red, 35.5%) across all three conditions.

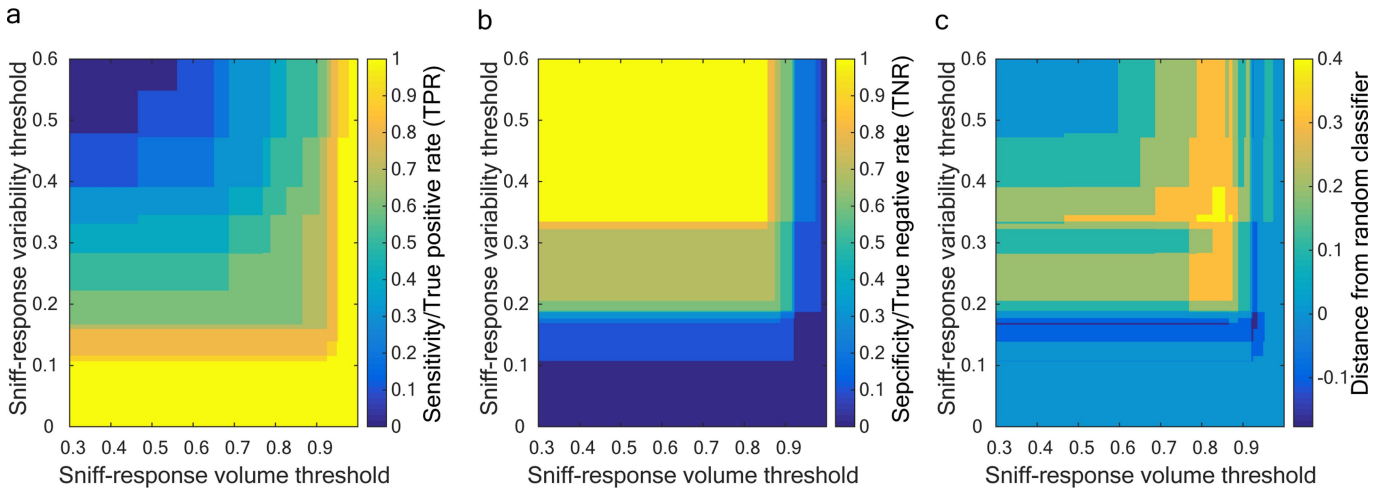


Extended Data Fig. 5 | See next page for caption.

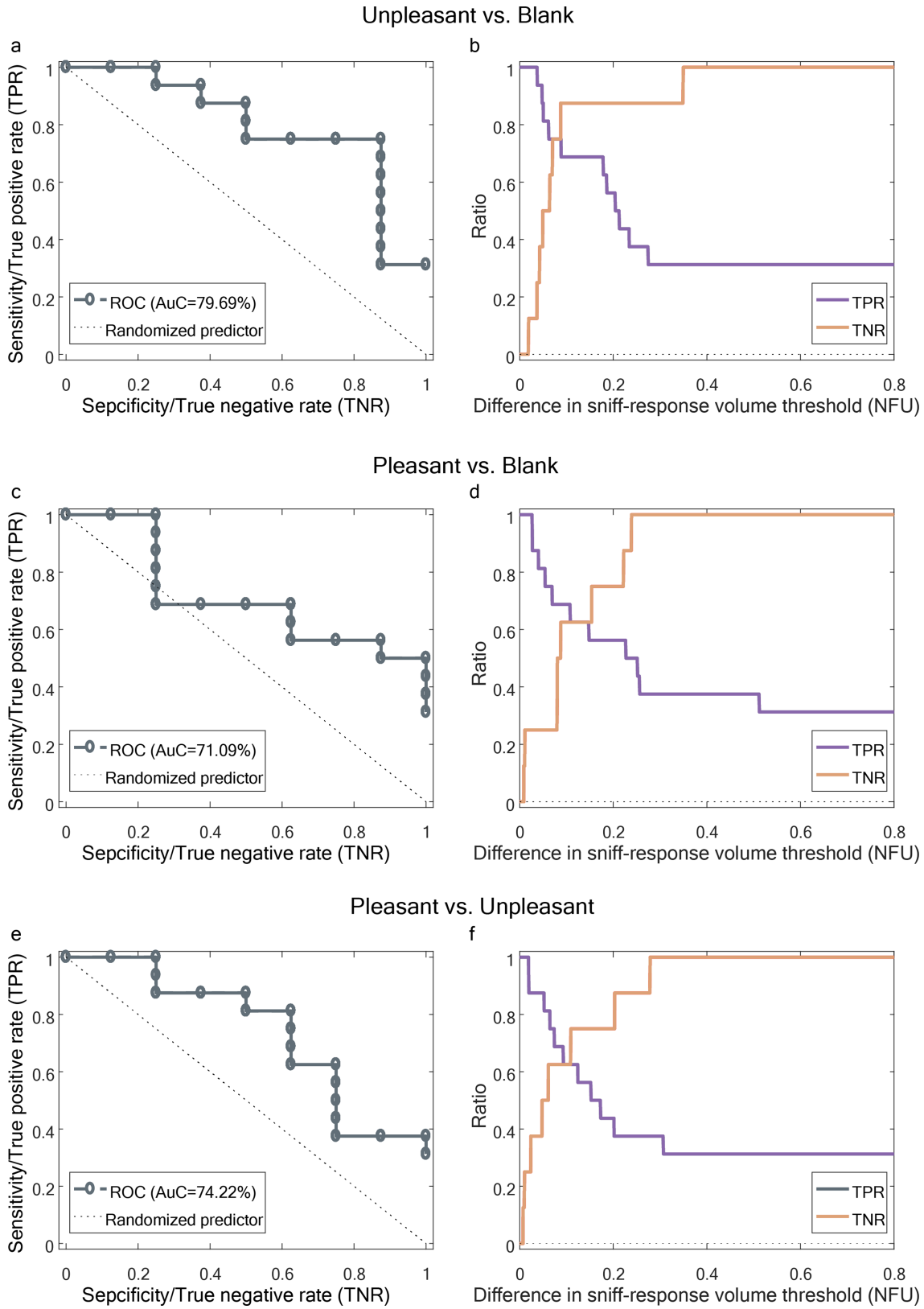
Extended Data Fig. 5 | The sniff response is prognostic for the recovery of consciousness and long-term survival in patients with DOC. Data are shown by patient rather than by session. The red lines denote the threshold of a sniff response (more than 15% change in magnitude and/or 0.35 s.d.). **a**–**c**, Each dot is a session with the strongest sniff response of a patient in a MCS ($n = 19$). **a**, Pleasant odorant. **b**, Unpleasant odorant. **c**, Blank. **d**, Percentage of patients in a MCS with sniff responses (white, 64.5%) and without sniff responses (red, 35.5%) across all three conditions. **e**–**g**, Each dot is a session with the strongest sniff response of a patient with UWS; empty dots represent patients who later recovered (transitioned to MCS; $n = 16$) and filled dots represent patients who did not recover and remain unconscious ($n = 8$). **e**, Pleasant odorant. **f**, Unpleasant odorant. **g**, Blank. **h**, Percentage of patients with UWS who later transitioned to MCS (left, recovered) and who remain unconscious (right, unrecovered) with sniff responses (white; recovered, 62.5%; unrecovered, 0%) and without sniff responses (red; recovered, 37.5%; unrecovered, 100%) across all three conditions. **i**–**k**, Each dot is a patient with DOC (MCS and UWS; $n = 43$); filled black dots represent patients who died during the study and coloured dots represent patients who survived during the study (mean ± s.d., 37.3 ± 14.1 months after brain injury). **i**, Pleasant odorant. **j**, Unpleasant odorant. **k**, Blank. **l**, Percentage of patients with DOC with sniff responses (left) who survived (white, 91.7%) and who are deceased (D) (red, 8.3%) and of patients with DOC without sniff responses who survived (white, 36.8%) and are deceased (red, 63.2%). **m**–**o**, Relation between the functional independence measure and normalized sniff volume. Each dot is a patient with UWS who survived during the study ($n = 12$). **m**, Pleasant odorant. **n**, Unpleasant odorant. **o**, Blank. r represents Spearman correlation.



Extended Data Fig. 6 | Dependence of the predictive value on sniff-response thresholds in patients with UWS. a, The receiver-operating characteristic (ROC) for a range of sniff-response volume thresholds for a sniff-response volume variability threshold of 0.35. **b**, The ROC for a range of sniff-response volume variability thresholds for a sniff-response volume threshold of 0.85 (15% reduction). **c**, Distance from a randomized predictor for a range of sniff-response volume thresholds for a sniff-response volume variability threshold of 0.35. **d**, Distance from randomized predictor for a range of sniff-response volume variability thresholds for a sniff-response volume threshold of 0.85. **e**, Area under the curve (AuC) for different sniff-response volume variability thresholds. **f**, Area under the curve for different sniff-response volume thresholds. $n = 24$. Further analyses are provided in the Supplementary Information.



Extended Data Fig. 7 | Dependence of the sensitivity and specificity on sniff-response thresholds. a, Sensitivity (true-positive rate (TPR)) for a range of sniff-response volume and sniff-response volume variability thresholds. **b**, Specificity (true-negative rate (TNR)) for a range of sniff-response volume and sniff-response volume variability thresholds. **c**, Distance from a randomized predictor for a range of sniff-response volume and sniff-response volume variability thresholds. Further analyses are provided in the Supplementary Information.



Extended Data Fig. 8 | Dependence of the predictive value on the relation between conditions. a, ROC for a range of differences in the sniff-response volume between the unpleasant odorant and the blank. **b**, The true-positive and true-negative rates for a range of differences in the sniff-response volume between the unpleasant odorant and the blank. **c**, ROC for a range of differences in the sniff-response volume between the pleasant odorant and the blank. **d**, The true-positive and true-negative rates for a range of differences in the sniff-response volume between the pleasant odorant and the blank. **e**, ROC for a range of differences in the sniff-response volume between the pleasant and unpleasant odorants. **f**, The true-positive and true-negative rates for a range of differences in the sniff-response volume between the pleasant and unpleasant odorants. $n = 24$. Further analyses are provided in the Supplementary Information.

Extended Data Table 1 | Patient information

Patient number	Consciousness state at enrolment	Age	Gender	Etiology	Time Since Onset (months)	Mean Inter-session Interval (weeks)	Number of included sessions	Number of sessions (all)
1	MCS	21	M	TBI	1	2.7	3	3
3	MCS	20	F	ABI	3.5	NA	1	2
4	MCS	59	M	TBI	4	3.5	3	3
5	UWS	26	F	TBI	7.5	3.1	3	3
6	UWS	27	M	TBI	3	2.9	5	5
7	UWS	23	M	TBI	1.5	1.7	9	9
8	UWS	38	M	ABI	2	1.3	12	12
9	UWS	19	M	ABI	2.5	1.4	10	10
11	UWS	30	M	CVA	3	1.7	12	13
12	MCS	44	M	ABI	3.5	NA	1	1
13	MCS	42	M	CVA	2.5	3.7	1	2
14	UWS	57	M	TBI	2.5	1.1	9	11
16	UWS	55	M	TBI	3.5	NA	1	3
17	UWS	42	M	TBI	5.5	NA	1	3
18	MCS	34	M	TBI	9.5	3.6	1	3
19	MCS	21	M	TBI	7	6.4	3	5
20	UWS	18	M	TBI	7	9.6	2	7
21	UWS	60	M	TBI	3.5	1.0	2	3
22	UWS	45	M	TBI	4	NA	1	4
23	MCS	44	M	CVA	2.5	2.7	5	5
24	UWS	66	M	TBI	6.5	2.4	3	4
25	MCS	38	M	TBI	2.5	3.2	3	3
26	MCS	70	M	TBI	1.5	NA	1	1
27	MCS	22	M	TBI	3	NA	1	3
28	UWS	49	M	CVA	5	2.6	2	4
30	MCS	22	M	TBI	1	NA	1	1
31	UWS	64	M	TBI	2	2.6	4	2
32	UWS	61	F	CVA	3.5	2.6	3	4
33	MCS	23	M	TBI	8	NA	1	1
34	MCS	19	M	TBI	2.5	2.5	6	8
35	MCS	75	F	Infection	2.5	NA	1	1
36	UWS	68	F	TBI	2.5	1.4	3	3
37	UWS	39	F	CVA	2	2.1	5	5
38	MCS	55	F	CVA	4.5	0.9	2	4
39	MCS	59	M	TBI	3	2.0	2	2
40	MCS	51	F	CVA	2.5	1.6	2	2
43	UWS	27	M	TBI	1	2.2	2	3
44	MCS	48	M	TBI	5.5	2.3	2	2
45	UWS	33	M	TBI	2	1.6	4	4
46	MCS	58	M	CVA	3.5	NA	1	3
47	UWS	48	M	ABI	6	NA	1	2
48	MCS	69	M	CVA	3	2.0	7	8
49	MCS	45	M	TBI	10	4.0	4	4
Total							**146**	**181**

ABI, anoxic brain injury; CVA, cerebrovascular accident; F, female; M, male; TBI, traumatic brain injury. $n = 43$. NA, not applicable, when only one included session was available for a patient, there was no inter-session interval.

Extended Data Table 2 | Damage to olfactory-related brain areas

Patient number	Etiology	Cribriform plate damage	Frontal lobe damage	Orbitofronal damage	Thalamic damage
1	TBI	No	No	No	No
3	ABI	No	No	No	No
4	TBI	No	Yes	Yes	No
5	TBI	No	No	No	No
6	TBI	Suspected	Yes	No	No
7	TBI	No	No	No	No
8	ABI	Suspected	No	No	No
9	ABI	No	No	No	No
11	CVA	No	Yes	Yes	No
12	ABI	No	No	No	No
13	CVA	No	Yes	Yes	Yes
14	TBI	No	No	No	Yes
16	TBI	Yes	Yes	No	Yes
17	TBI	No	No	No	No
18	TBI	Suspected	Yes	Yes	No
19	TBI	Suspected	No	No	No
20	TBI	No	Yes	No	No
21	TBI	No	No	No	Yes
22	TBI	No	Yes	Yes	No
23	CVA	No	No	No	No
24	TBI	No	No	No	No
25	TBI	No	No	No	No
26	TBI	No	Yes	No	No
27	TBI	No	Yes	Yes	No
28	CVA	No	No	No	No
30	TBI	No	Yes	Yes	No
31	TBI	No	No	No	No
32	CVA	No	Yes	No	No
33	TBI	Suspected	Yes	Yes	Yes
34	TBI	No	Yes	No	No
35	Infection	No	No	No	No
36	TBI	No	Yes	No	Yes
37	CVA	No	Yes	No	No
38	CVA	No	Yes	Yes	Yes
39	TBI	No	Yes	No	No
40	CVA	No	No	No	Yes
43	TBI	No	No	No	Yes
44	TBI	No	Yes	Yes	No
45	TBI	No	No	No	No
46	CVA	No	No	No	Yes
47	ABI	No	No	No	No
48	CVA	No	No	No	No
49	TBI	No	Yes	No	No

Readings of structural imaging (computed tomography and magnetic resonance imaging), which were mostly conducted directly after injury ($n = 43$). If traumatic brain injuries were not equally distributed across the patient subgroups, this may have biased the outcome. To address this, we compared the proportion of traumatic brain injuries in each patient group (patients in a MCS, patients who transitioned between MCS and UWS, and patients with UWS) and observed no differences (11 out of 19 patients in a MCS, 10 out of 16 transitioning patients, and 6 out of 8 UWS; $\chi^2 = 0.71$, $P = 0.7$). Moreover, we repeated the analysis excluding the 27 patients with traumatic brain injuries. The remaining aetiologies were anoxic brain injuries ($n = 5$), cerebrovascular accidents ($n = 10$) and infection ($n = 1$). Repeating the analysis with only 37% of the original sample size, similar results were obtained. At the first sniff after odorant presentation, MCS sniff volume was significantly smaller than UWS sniff volume for both pleasant (MCS, mean ± s.d. = 0.88 ± 0.22 NFU, median = 0.91 NFU; UWS, mean ± s.d. = 0.97 ± 0.15 NFU, median = 1.0 NFU; $Z = 2.2$, $P = 0.027$) and unpleasant odorants (MCS, mean ± s.d. = 0.84 ± 0.20 NFU, median = 0.89 NFU; UWS, mean ± s.d. = 0.99 ± 0.13 NFU, median = 0.99 NFU; $Z = 3.26$, $P = 0.001$). As for the cognitively driven sniff response, we observed only a weak trend in this limited sample (MCS, mean ± s.d. = 0.965 ± 0.12 NFU, median = 0.99 NFU; UWS, mean ± s.d. = 1.0 ± 0.18 NFU, median = 1.0 NFU; $Z = 1.1$, $P = 0.274$).

nature research

Corresponding author(s): Anat Arzi and Noam Sobel

Last updated by author(s): Feb 5, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Labchart v7.2.5, ADInstruments.
Data analysis	Matlab 2012b, R version 3.4.3

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw nasal airflow data collected in this study, with its associated behavioural assessments, will be posted for download at https://www.weizmann.ac.il/neurobiology/worg/materials.html after acceptance for publication.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	50 patients with 190 sessions were collected. We generate the sample size consistent with published studies in this field (Nigri et al., 2015, EJN)
Data exclusions	Data exclusion was based on nasal respiration. Out of 50 patients and 190 sessions, 43 patients and 146 sessions were included. Total of 41 olfactory testing sessions were excluded due to lack of stable nasal respiration and 3 sessions were excluded due to sniff volumed extending 3 SD. Seven patients were excluded due lack of a single session with stable nasal respiration. The rational behind session exclusion is insufficient data collection. The rational behind trials exclusion was unreliability due to noise. The number of trials required for inclusion was based on previous studies (Arzi et al., 2012 Nat. Neuro; Arzi et al., 2014, J. Neuro.). Detailed inclusion/exclusion criteria can be found in the methods section
Replication	To verify the reproducibility of the findings we conducted several control analyses, verifying that the findings are driven by olfactory sniff-responses and not by odor trigeminality , nor by changes in nasal respiration due to tracheostomy. Replication of sniff-response in MCS but not in VS/UWS using pure olfactory odorants was successful. In addition, we conducted effect size analyses to provide information on the strength of evidence supporting the findings.
Randomization	Patients were allocated to VS/UWS or MCS group based on behavioural assessments of their consciousness state on the day of the olfactory sniff-response test. Therefore, randomization between VS/UWS and MCS is not relevant.
Blinding	On the first olfactory testing session of each patients the experimenters were unaware of the patients' state of consciousness and conducted an independent behavioral evaluation. As patients state of consciousness fluctuate data acquisition of the behavioral state was done in each session by the same experimenter and therefore blinding beyond the first session was not applicable. The analysis of sniff-responses is automated and blind to the patients' state of consciousness, recovery and survival.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Disorders of consciousness (DoC) patients
Recruitment	Patients were recruited by the Loewenstein rehabilitation hospital physicians. The legal guardian of any patient arriving to the intensive care and rehabilitation of consciousness unit in the hospital were asked if they are willing to participate in the study. The physicians recruiting the patients were not involved in data collection in order to minimize any potential bias in recruitment.
Ethics oversight	The study was approved by the ethics committee of Loewenstein rehabilitation hospital. Written informed consent was obtained from the patients' legal guardian.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

The mutational constraint spectrum quantified from variation in 141,456 humans

https://doi.org/10.1038/s41586-020-2308-7

Received: 27 January 2019

Accepted: 26 March 2020

Published online: 27 May 2020

Open access

Check for updates

Konrad J. Karczewski[1,2]✉, Laurent C. Francioli[1,2], Grace Tiao[1,2], Beryl B. Cummings[1,2,3], Jessica Alföldi[1,2], Qingbo Wang[1,2,4], Ryan L. Collins[1,4,5], Kristen M. Laricchia[1,2], Andrea Ganna[1,2,6], Daniel P. Birnbaum[1,2], Laura D. Gauthier[7], Harrison Brand[1,5], Matthew Solomonson[1,2], Nicholas A. Watts[1,2], Daniel Rhodes[8], Moriel Singer-Berk[1,2], Eleina M. England[1,2], Eleanor G. Seaby[1,2], Jack A. Kosmicki[1,2,4], Raymond K. Walters[1,2,9], Katherine Tashman[1,2,9], Yossi Farjoun[7], Eric Banks[7], Timothy Poterba[1,2,9], Arcturus Wang[1,2,9], Cotton Seed[1,2,9], Nicola Whiffin[1,2,10,11], Jessica X. Chong[12], Kaitlin E. Samocha[13], Emma Pierce-Hoffman[1,2], Zachary Zappala[1,2,14], Anne H. O'Donnell-Luria[1,2,15,16], Eric Vallabh Minikel[1], Ben Weisburd[7], Monkol Lek[17], James S. Ware[1,10,11], Christopher Vittal[2,9], Irina M. Armean[1,2], Louis Bergelson[7], Kristian Cibulskis[7], Kristen M. Connolly[18], Miguel Covarrubias[7], Stacey Donnelly[1], Steven Ferriera[18], Stacey Gabriel[18], Jeff Gentry[7], Namrata Gupta[1,18], Thibault Jeandet[7], Diane Kaplan[7], Christopher Llanwarne[7], Ruchi Munshi[7], Sam Novod[7], Nikelle Petrillo[7], David Roazen[7], Valentin Ruano-Rubio[7], Andrea Saltzman[1], Molly Schleicher[1], Jose Soto[7], Kathleen Tibbetts[7], Charlotte Tolonen[7], Gordon Wade[7], Michael E. Talkowski[1,5,19], Genome Aggregation Database Consortium*, Benjamin M. Neale[1,2,9], Mark J. Daly[1,2,6,9] & Daniel G. MacArthur[1,2,149,150]✉

Genetic variants that inactivate protein-coding genes are a powerful source of information about the phenotypic consequences of gene disruption: genes that are crucial for the function of an organism will be depleted of such variants in natural populations, whereas non-essential genes will tolerate their accumulation. However, predicted loss-of-function variants are enriched for annotation errors, and tend to be found at extremely low frequencies, so their analysis requires careful variant annotation and very large sample sizes[1]. Here we describe the aggregation of 125,748 exomes and 15,708 genomes from human sequencing studies into the Genome Aggregation Database (gnomAD). We identify 443,769 high-confidence predicted loss-of-function variants in this cohort after filtering for artefacts caused by sequencing and annotation errors. Using an improved model of human mutation rates, we classify human protein-coding genes along a spectrum that represents tolerance to inactivation, validate this classification using data from model organisms and engineered human cells, and show that it can be used to improve the power of gene discovery for both common and rare diseases.

The physiological function of most genes in the human genome remains unknown. In biology, as in many engineering and scientific fields, breaking the individual components of a complex system can provide valuable insight into the structure and behaviour of that system. For the discovery of gene function, a common approach is to introduce disruptive mutations into genes and determine their effects on cellular and physiological phenotypes in mutant organisms or cell lines[2]. Such studies have yielded valuable insight into eukaryotic physiology and have guided the design of therapeutic agents[3]. However, although studies in model organisms and human cell lines have been crucial in deciphering the function of many human genes, they remain imperfect proxies for human physiology.

Obvious ethical and technical constraints prevent the large-scale engineering of loss-of-function mutations in humans. However, recent exome and genome sequencing projects have revealed a surprisingly high burden of natural pLoF variation in the human population,

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [3]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. [4]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA. [5]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [6]Institute for Molecular Medicine Finland, Helsinki, Finland. [7]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Centre for Translational Bioinformatics, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London and Barts Health NHS Trust, London, UK. [9]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [10]National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. [11]Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London, UK. [12]Department of Pediatrics, University of Washington, Seattle, WA, USA. [13]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [14]Vertex Pharmaceuticals Inc, Boston, MA, USA. [15]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [16]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [17]Department of Genetics, Yale School of Medicine, New Haven, CT, USA. [18]Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [19]Department of Neurology, Harvard Medical School, Boston, MA, USA. [149]Present address: Centre for Population Genomics, Garvan Institute of Medical Research, and UNSW Sydney, Sydney, New South Wales, Australia. [150]Present address: Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: konradk@broadinstitute.org; d.macarthur@garvan.org.au

including stop-gained, essential splice, and frameshift variants[1,4], which can serve as natural models for inactivation of human genes. Such variants have already revealed much about human biology and disease mechanisms, through many decades of study of the genetic basis of severe Mendelian diseases[5], most of which are driven by disruptive variants in either the heterozygous or homozygous state. These variants have also proved valuable in identifying potential therapeutic targets: confirmed LoF variants in the *PCSK9* gene have been causally linked to low levels of low-density lipoprotein cholesterol[6], and have ultimately led to the development of several inhibitors of PCSK9 that are now in clinical use for the reduction of cardiovascular disease risk. A systematic catalogue of pLoF variants in humans and the classification of genes along a spectrum of tolerance to inactivation would provide a valuable resource for medical genetics, identifying candidate disease-causing mutations, potential therapeutic targets, and windows into the normal function of many currently uncharacterized human genes.

Several challenges arise when assessing LoF variants at scale. LoF variants are on average deleterious, and are thus typically maintained at very low frequencies in the human population. Systematic genome-wide discovery of these variants requires whole-exome or whole-genome sequencing of very large numbers of samples. In addition, LoF variants are enriched for false positives compared with synonymous or other benign variants, including mapping, genotyping (including somatic variation), and particularly, annotation errors[1], and careful filtering is required to remove such artefacts.

Population surveys of coding variation enable the evaluation of the strength of natural selection at a gene or region level. As natural selection purges deleterious variants from human populations, methods to detect selection have modelled the reduction in variation (constraint)[7] or shift in the allele frequency distribution[8], compared to an expectation. For analyses of selection on coding variation, synonymous variation provides a convenient baseline, controlling for other potential population genetic forces that may influence the amount of variation as well as technical features of the local sequence. A model of constraint was previously applied to define a set of 3,230 genes with a high probability of intolerance to heterozygous pLoF variation (pLI)[4] and estimated the selection coefficient for variants in these genes[9]. However, the ability to comprehensively characterize the degree of selection against pLoF variants is particularly limited, as for small genes, the expected number of mutations is still very low, even for samples of up to 60,000 individuals[4,10]. Furthermore, the previous dichotomization of pLI, although convenient for the characterization of a set of genes, disguises variability in the degree of selective pressure against a given class of variation and overlooks more subtle levels of intolerance to pLoF variation. With larger sample sizes, a more accurate quantitative measure of selective pressure is possible.

Here, we describe the detection of pLoF variants in a cohort of 125,748 individuals with whole-exome sequence data and 15,708 individuals with whole-genome sequence data, as part of the Genome Aggregation Database (gnomAD; https://gnomad.broadinstitute.org), the successor to the Exome Aggregation Consortium (ExAC). We develop a continuous measure of intolerance to pLoF variation, which places each gene on a spectrum of LoF intolerance. We validate this metric by comparing its distribution to several orthogonal indicators of constraint, including the incidence of structural variation and the essentiality of genes as measured using mouse gene knockout experiments and cellular inactivation assays. Finally, we demonstrate that this metric improves the interpretation of genetic variants that influence rare disease and provides insight into common disease biology. These analyses provide, to our knowledge, the most comprehensive catalogue so far of the sensitivity of human genes to disruption.

In a series of accompanying manuscripts, other complementary analyses of this dataset are described. Using an overlapping set of 14,237 whole genomes, the discovery and characterization of a wide variety of structural variants (large deletions, duplications, insertions, or other rearrangements of DNA) is reported[11]. The value of pLoF variants for the discovery and validation of therapeutic drug targets is explored[12], and a case study of the use of these variants from gnomAD and other large reference datasets is provided to validate the safety of inhibition of LRRK2—a candidate therapeutic target for Parkinson's disease[13]. By combining the gnomAD dataset with a large collection of RNA sequencing data from adult human tissues[14], the value of tissue expression data in the interpretation of genetic variation across a range of human diseases is reported[15]. Finally, the effect of two understudied classes of human variation—multi-nucleotide variants[16] and variants that create or disrupt open-reading frames in the 5′ untranslated region of human genes—is characterized and investigated[17].

A high-quality catalogue of variation

We aggregated whole-exome sequencing data from 199,558 individuals and whole-genome sequencing data from 20,314 individuals. These data were obtained primarily from case–control studies of common adult-onset diseases, including cardiovascular disease, type 2 diabetes and psychiatric disorders. Each dataset, totalling more than 1.3 and 1.6 petabytes of raw sequencing data, respectively, was uniformly processed, joint variant calling was performed on each dataset using a standardized BWA-Picard-GATK pipeline[18], and all data processing and analysis was performed using Hail[19]. We performed stringent sample quality control (Extended Data Fig. 1), removing samples with lower sequencing quality by a variety of metrics, samples from second-degree or closer related individuals across both data types, samples with inadequate consent for the release of aggregate data, and samples from individuals known to have a severe childhood-onset disease as well as their first-degree relatives. The final gnomAD release contains genetic variation from 125,748 exomes and 15,708 genomes from unique unrelated individuals with high-quality sequence data, spanning 6 global and 8 sub-continental ancestries (Fig. 1a, b), which we have made publicly available at https://gnomad.broadinstitute.org. We also provide subsets of the gnomAD datasets, which exclude individuals who are cases in case–control studies, or who are cases of a few particular disease types such as cancer and neurological disorders, or who are also aggregated in the Bravo TOPMed variant browser (https://bravo.sph.umich.edu).

Among these individuals, we discovered 17.2 million and 261.9 million variants in the exome and genome datasets, respectively; these variants were filtered using a custom random forest process (Supplementary Information) to 14.9 million and 229.9 million high-quality variants. Comparing our variant calls in two samples for which we had independent gold-standard variant calls, we found that our filtering achieves very high precision (more than 99% for single nucleotide variants (SNVs), over 98.5% for indels in both exomes and genomes) and recall (over 90% for SNVs and more than 82% for indels for both exomes and genomes) at the single sample level (Extended Data Fig. 2). In addition, we leveraged data from 4,568 and 212 trios included in our exome and genome call-sets, respectively, to assess the quality of our rare variants. We found that our model retains over 97.8% of the transmitted singletons (singletons in the unrelated individuals that are transmitted to an offspring) on chromosome 20 (which was not used for model training) (Extended Data Fig. 3a–d). In addition, the number of putative de novo calls after filtering are in line with expectations[20] (Extended Data Fig. 3e–h), and our model had a recall of 97.3% for de novo SNVs and 98% for de novo indels based on 375 independently validated de novo variants in our whole-exome trios (295 SNVs and 80 indels) (Extended Data Fig. 3i, j). Altogether, these results indicate that our filtering strategy produced a call-set with high precision and recall for both common and rare variants.

These variants reflect the expected patterns based on mutation and selection: we observe 84.9% of all possible consistently methylated CpG-to-TpG transitions that would create synonymous variants in the human exome (Supplementary Table 14), which indicates that at this

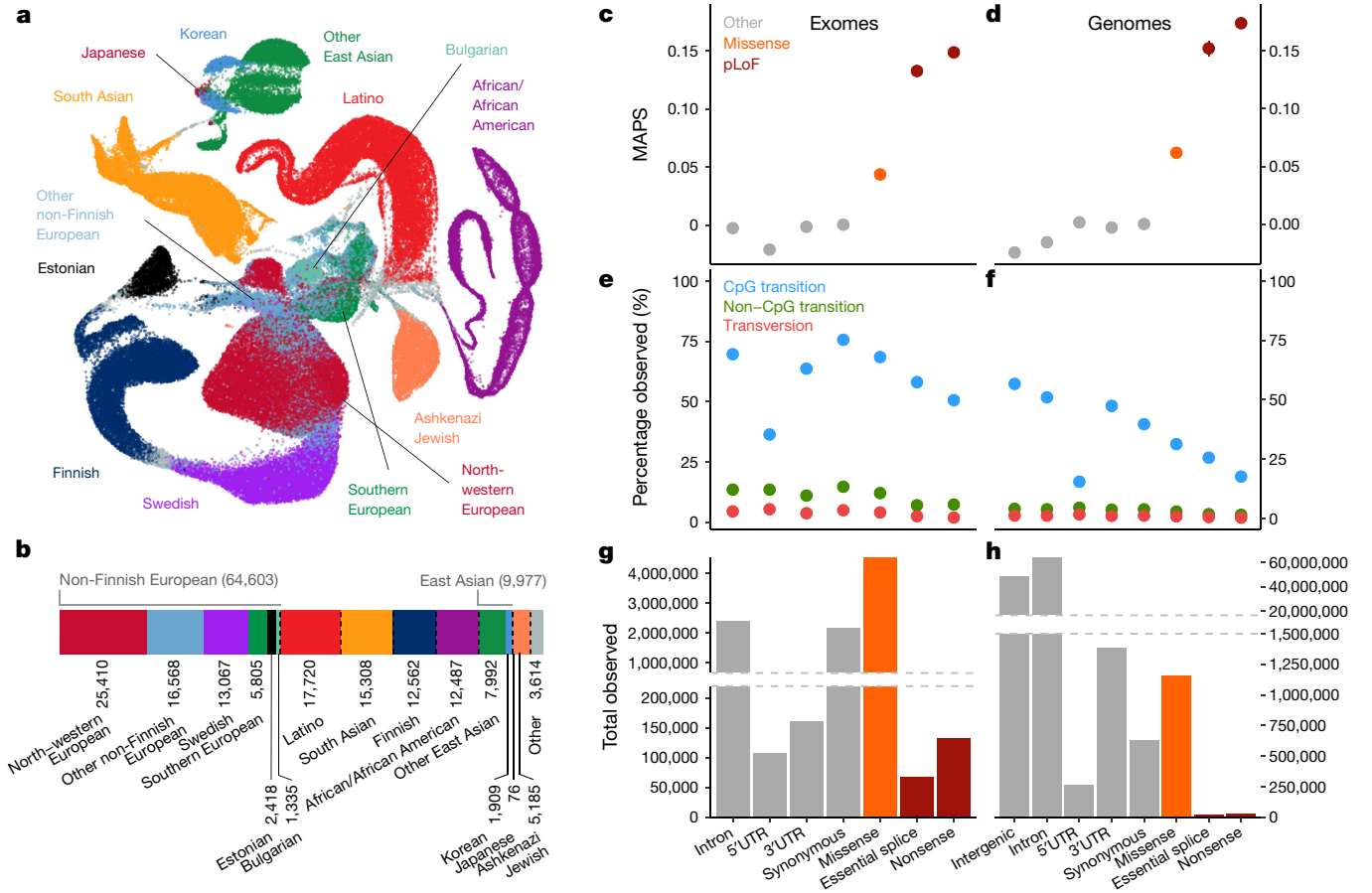


Fig. 1 | Aggregation of 141,456 exome and genome sequences. a, Uniform manifold approximation and projection (UMAP)[46,47] plot depicting the ancestral diversity of all individuals in gnomAD, using ten principal components. Note that long-range distances in the UMAP space are not a proxy for genetic distance. **b**, The number of individuals by population and subpopulation in the gnomAD database. Colours representing populations in **a** and **b** are consistent. **c**, **d**, The mutability-adjusted proportion of singletons[4] (MAPS) is shown across functional categories for SNVs in exomes (**c**; *x* axis shared with **e** and **g**) and genomes (**d**; *x* axis shared with **f** and **h**). Higher values indicate an enrichment of lower frequency variants, which suggests increased deleteriousness. **e**, **f**, The proportion of possible variants observed for each functional class for each mutational type for exomes (**e**) and genomes (**f**). CpG transitions are more saturated, except where selection (for example, pLoFs) or hypomethylation (5′ untranslated region) decreases the number of observations. **g**, **h**, The total number of variants observed in each functional class for exomes (**g**) and genomes (**h**). Error bars in **c**–**f** represent 95% confidence intervals (note that in some cases these are fully contained within the plotted point).

sample size, we are beginning to approach mutational saturation of this highly mutable and weakly negatively selected variant class. However, we only observe 52% of methylated CpG stop-gained variants, which illustrates the action of natural selection removing a substantial fraction of gene-disrupting variants from the population (Fig. 1c–h). Across all mutational contexts, only 11.5% and 3.7% of the possible synonymous and stop-gained variants, respectively, are observed in the exome dataset, which indicates that current sample sizes remain far from capturing complete mutational saturation of the human exome (Extended Data Fig. 4).

Identifying loss-of-function variants

Some LoF variants will result in embryonic lethality in humans in a heterozygous state, whereas others are benign even at homozygosity, with a wide spectrum of effects in between. Throughout this manuscript, we define pLoF variants to be those that introduce a premature stop (stop-gained), shift-reported transcriptional frame (frameshift), or alter the two essential splice-site nucleotides immediately to the left and right of each exon (splice) found in protein-coding transcripts, and ascertain their presence in the cohort of 125,748 individuals with exome sequence data. As these variants are enriched for annotation artefacts[1], we developed the loss-of-function transcript effect estimator (LOFTEE) package, which applies stringent filtering criteria from first principles (such as removing terminal truncation variants, as well as rescued splice variants, that are predicted to escape nonsense-mediated decay) to pLoF variants annotated by the variant effect predictor (Extended Data Fig. 5a). Despite not using frequency information, we find that this method disproportionately removes pLoF variants that are common in the population, which are known to be enriched for annotation errors[1], while retaining rare, probable deleterious variations, as well as reported pathogenic variation (Fig. 2a). LOFTEE distinguishes high-confidence pLoF variants from annotation artefacts, and identifies a set of putative splice variants outside the essential splice site. The filtering strategy of LOFTEE is conservative in the interest of increasing specificity, filtering some potentially functional variants that display a frequency spectrum consistent with that of missense variation (Fig. 2b). Applying LOFTEE v1.0, we discover 443,769 high-confidence pLoF variants, of which 413,097 fall on the canonical transcripts of 16,694 genes. The number of pLoF variants per individual is consistent with previous reports[1], and is highly dependent on the frequency filters chosen (Supplementary Table 17).

Aggregating across variants, we created a gene-level pLoF frequency metric to estimate the proportion of haplotypes that contain an inactive

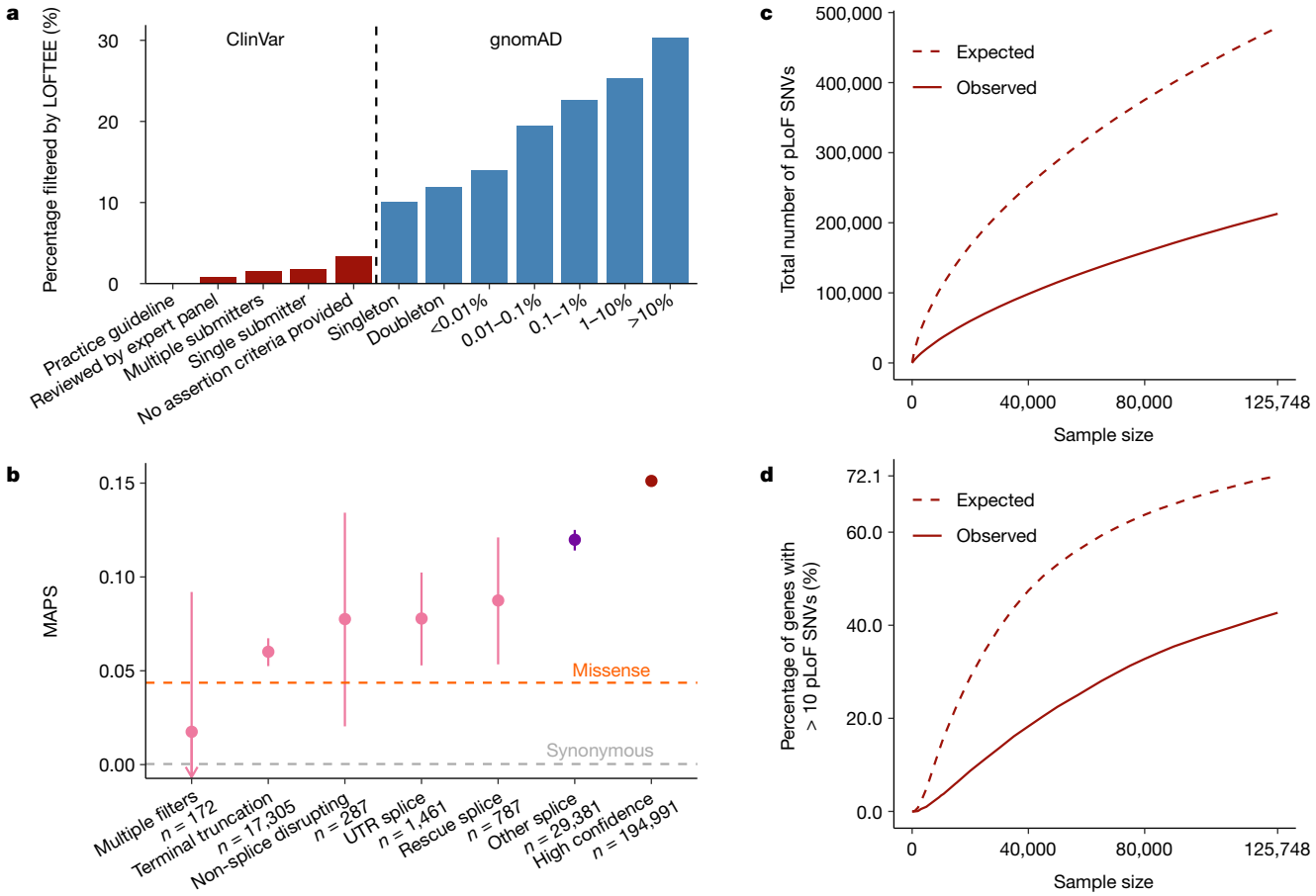


Fig. 2 | Generating a high-confidence set of pLoF variants. a, The percentage of variants filtered by LOFTEE grouped by ClinVar status and gnomAD frequency. Despite not using frequency information, LOFTEE removes a larger proportion of common variants, and a very low proportion of reported disease-causing variation. **b**, MAPS (see Fig. 1c, d) is shown by LOFTEE designation and filter. Variants filtered out by LOFTEE exhibit frequency spectra that are similar to those of missense variants; predicted splice variants outside the essential splice site are more rare, and high-confidence variants are very likely to be singletons. Only SNVs with at least 80% call rate are included here. Error bars represent 95% confidence intervals. **c**, **d**, The total number of pLoF variants (**c**), and proportion of genes with more than ten pLoF variants (**d**) observed and expected (in the absence of selection) as a function of sample size (downsampled from gnomAD). Selection reduces the number of variants observed, and variant discovery approximately follows a square-root relationship with the number of samples. At current sample sizes, we would expect to identify more than 10 pLoF variants for 72.1% of genes in the absence of selection.

copy of each gene. We find that 1,555 genes have an aggregate pLoF frequency of at least 0.1% across all individuals in the dataset (Extended Data Fig. 5c), and 3,270 genes have an aggregate pLoF frequency of at least 0.1% in any one population. Furthermore, we characterized the landscape of genic tolerance to homozygous inactivation, identifying 4,332 pLoF variants that are homozygous in at least one individual. Given the rarity of true homozygous LoF variants, we expected substantial enrichment of such variants for sequencing and annotation errors, and we subjected this set to additional filtering and deep manual curation before defining a set of 1,815 genes (2,636 high-confidence variants) that are likely to be tolerant to biallelic inactivation (Supplementary Data 7).

The LoF intolerance of human genes

Just as a preponderance of pLoF variants is useful for identifying LoF-tolerant genes, we can conversely characterize the intolerance of a gene to inactivation by identifying marked depletions of predicted LoF variation[4,7]. Here, we present a refined mutational model, which incorporates methylation, base-level coverage correction, and LOFTEE (Supplementary Information, Extended Data Fig. 6), to predict expected levels of variation under neutrality. Under this updated model, the variation in the number of synonymous variants observed is accurately captured ($r = 0.979$). We then applied this method to detect depletion of pLoF variation by comparing the number of observed pLoF variants against our expectation in the gnomAD exome data from 125,748 individuals—more than doubling the sample size of ExAC, the previously largest exome collection[4]. For this dataset, we computed a median of 17.9 expected pLoF variants per gene (Fig. 2c) and found that 72.1% of genes have more than 10 pLoF variants (powered to be classified into the most constrained genes) (Supplementary Information) expected on the canonical transcript (Fig. 2d), an increase from 13.2% and 62.8%, respectively, in ExAC.

The smaller sample size in ExAC required a transformation of the observed and expected values for the number of pLoF variants in each gene into the pLI: this metric estimates the probability that a gene falls into the class of LoF-haploinsufficient genes (approximately 10% observed/expected variation) and is ideally used as a dichotomous metric (producing 3,230 genes with pLI > 0.9). Here, our refined model and substantially increased sample size enabled us to directly assess the degree of intolerance to pLoF variation in each gene using the continuous metric of the observed/expected ratio and to estimate a confidence interval around the ratio. We find that the median observed/expected ratio is 48%, which indicates that, as noted previously, most genes

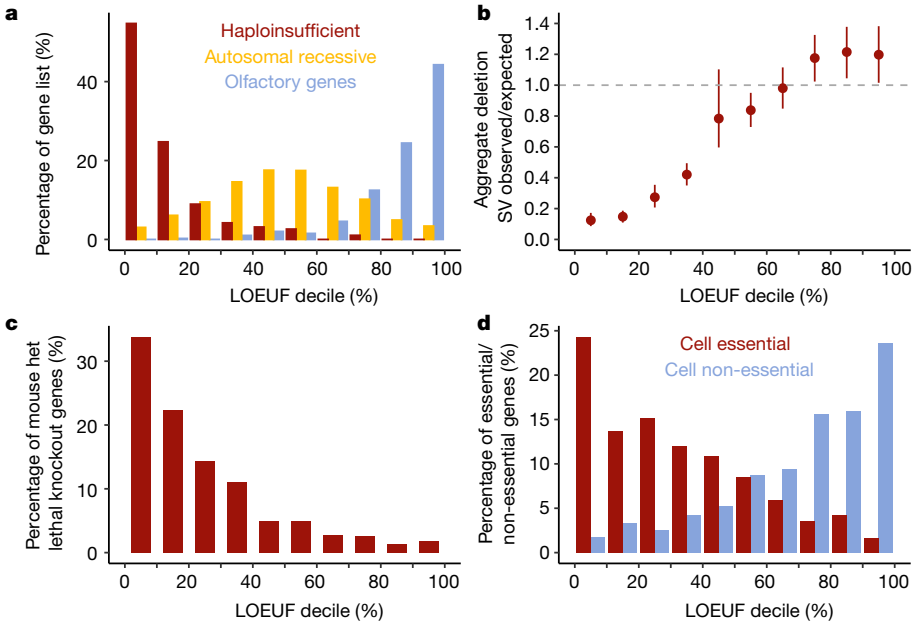


Fig. 3 | The functional spectrum of pLoF impact. a, The percentage of genes in a set of curated gene lists represented in each LOEUF decile. Haploinsufficient genes are enriched among the most constrained genes, whereas recessive genes are spread in the middle of the distribution, and olfactory receptor genes are largely unconstrained. **b**, The occurrence of 6,735 rare LoF deletion structural variants (SVs) is correlated with LOEUF (computed from SNVs; linear regression $r = 0.13$; $P = 9.8 \times 10^{-68}$). Error bars represent 95% confidence intervals from bootstrapping. **c**, **d**, Constrained genes are more likely to be lethal when heterozygously inactivated in mouse and cause cellular lethality when disrupted in human cells (**c**), whereas unconstrained genes are more likely to be tolerant of disruption in cellular models (**d**). For all panels, more constrained genes are shown on the left.

exhibit at least moderate selection against pLoF variation, and that the distribution of the observed/expected ratio is not dichotomous, but continuous (Extended Data Fig. 7a). For downstream analyses, unless otherwise specified, we use the 90% upper bound of this confidence interval, which we term the loss-of-function observed/expected upper bound fraction (LOEUF) (Extended Data Fig. 7b, c), and bin 19,197 genes into deciles of approximately 1,920 genes each. At current sample sizes, this metric enables the quantitative assessment of constraint with a built-in confidence value, and distinguishes small genes (for example, those with observed = 0, expected = 2; LOEUF = 1.34) from large genes (for example, observed = 0, expected = 100; LOEUF = 0.03), while retaining the continuous properties of the direct estimate of the ratio (Supplementary Information). At one extreme of the distribution, we observe genes with a very strong depletion of pLoF variation (first LOEUF decile aggregate observed/expected approximately 6%) (Extended Data Fig. 7e), including genes previously characterized as high pLI (Extended Data Fig. 7f). By contrast, we find unconstrained genes that are relatively tolerant of inactivation, including many that contain homozygous pLoF variants (Extended Data Fig. 7g).

We note that the use of the upper bound means that LOEUF is a conservative metric in one direction: genes with low LOEUF scores are confidently depleted for pLoF variation, whereas genes with high LOEUF scores are a mixture of genes without depletion, and genes that are too small to obtain a precise estimate of the observed/expected ratio. In general, however, the scale of gnomAD means that gene length is rarely a substantive confounder for the analyses described here, and all downstream analyses are adjusted for the length of the coding sequence or filtered to genes with at least ten expected pLoFs (Supplementary Information).

Validation of the LoF-intolerance score

The LOEUF metric allows us to place each gene along a continuous spectrum of tolerance to inactivation. We examined the correlation of this metric with several independent measures of genic sensitivity to disruption. First, we found that LOEUF is consistent with the expected behaviour of well-established gene sets: known haploinsufficient genes are strongly depleted of pLoF variation, whereas olfactory receptors are relatively unconstrained, and genes with a known autosomal recessive mechanism, for which selection against heterozygous disruptive variants tends to be present but weak[9], fall in the middle of the distribution (Fig. 3a). In addition, LOEUF is positively correlated with the occurrence of 6,735 rare autosomal deletion structural variants overlapping protein-coding exons identified in a subset of 6,749 individuals with whole-genome sequencing data in this manuscript[11] ($r = 0.13$; $P = 9.8 \times 10^{-68}$) (Fig. 3b).

This constraint metric also correlates with results in model systems: in 389 genes with orthologues that are embryonically lethal after heterozygous deletion in mouse[21,22], we find a lower LOEUF score (mean = 0.488), compared with the remaining 18,808 genes (mean = 0.962; t-test $P = 10^{-78}$) (Fig. 3c). Similarly, the 678 genes that are essential for human cell viability as characterized by CRISPR screens[23] are also depleted for pLoF variation (mean LOEUF = 0.63) in the general population compared to background (18,519 genes with mean LOEUF = 0.964; t-test $P = 9 \times 10^{-71}$), whereas the 777 non-essential genes are more likely to be unconstrained (mean LOEUF = 1.34, compared to remaining 18,420 genes with mean LOEUF = 0.936; t-test $P = 3 \times 10^{-92}$) (Fig. 3d).

Biological properties of constraint

We investigated the properties of genes and transcripts as a function of their tolerance to pLoF variation (LOEUF). First, we found that LOEUF correlates with the degree of connection of a gene in protein-interaction networks ($r = -0.14$; $P = 1.7 \times 10^{-51}$ after adjusting for gene length) (Fig. 4a) and functional characterization (Extended Data Fig. 8a). In addition, constrained genes are more likely to be ubiquitously expressed across 38 tissues in the Genotype-Tissue Expression

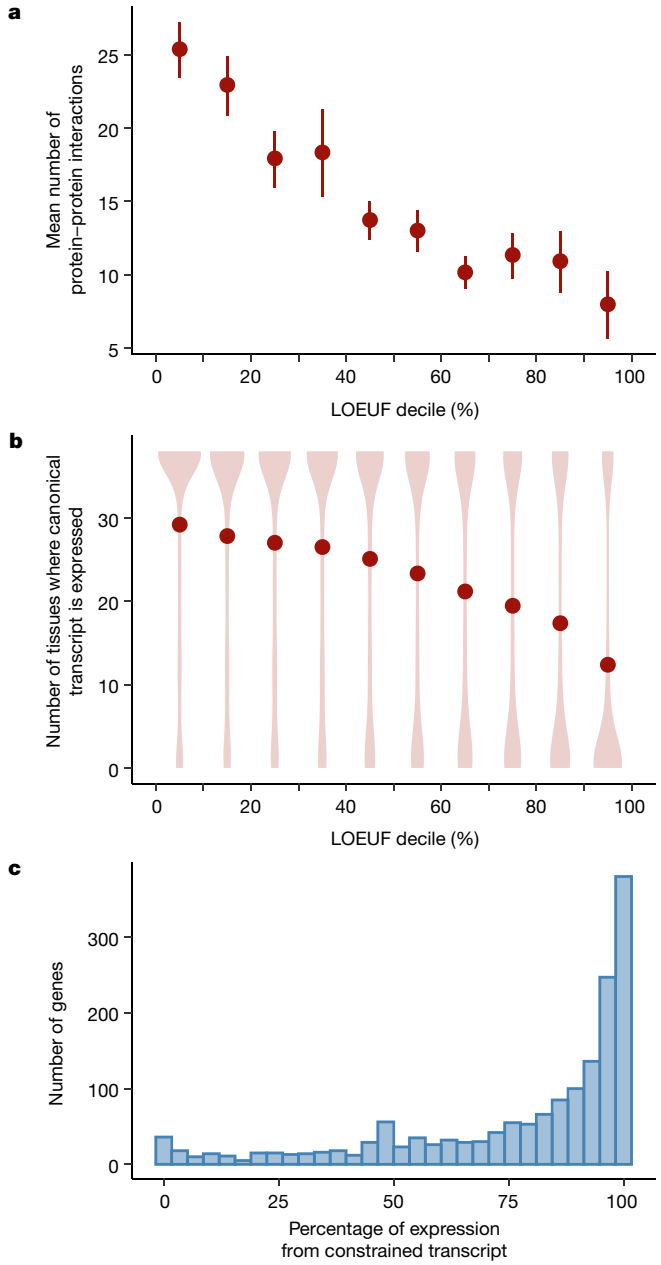


Fig. 4 | Biological properties of constrained genes and transcripts. a, The mean number of protein–protein interactions is plotted as a function of LOEUF decile: more constrained genes have more interaction partners (LOEUF linear regression $r = -0.14$; $P = 1.7 \times 10^{-51}$). Error bars correspond to 95% confidence intervals. **b**, The number of tissues where a gene is expressed (transcripts per million > 0.3), binned by LOEUF decile, is shown as a violin plot with the mean number overlaid as points: more constrained genes are more likely to be expressed in several tissues (LOEUF linear regression $r = -0.31$; $P < 1 \times 10^{-100}$). **c**, For 1,740 genes in which there exists at least one constrained and one unconstrained transcript, the proportion of expression derived from the constrained transcript is plotted as a histogram.

(GTEx) project (Fig. 4b) (LOEUF $r = -0.31$; $P < 1 \times 10^{-100}$) and have higher expression on average (LOEUF $\rho = -0.28$; $P < 1 \times 10^{-100}$), consistent with previous results[4]. Although most results in this study are reported at the gene level, we have also extended our framework to compute LOEUF for all protein-coding transcripts, allowing us to explore the extent of differential constraint of transcripts within a given gene. In cases in which a gene contained transcripts with varying levels of constraint, we found that transcripts in the first LOEUF decile were more likely to be expressed across tissues than others in the same gene ($n = 1{,}740$ genes), even when adjusted for transcript length (Fig. 4c) (constrained transcripts are on average 6.34 transcripts per million higher; $P = 2.2 \times 10^{-14}$). Furthermore, we found that the most constrained transcript for each gene was typically the most highly expressed transcript in tissues with disease relevance[24] (Extended Data Fig. 8c), which supports the need for transcript-based variant interpretation, as explored in more depth in an accompanying manuscript[15].

Finally, we investigated potential differences in LOEUF across human populations, restricting to the same sample size across all populations to remove bias due to differential power for variant discovery. As the smallest population in our exome dataset (African/African American) has only 8,128 individuals, our ability to detect constraint against pLoF variants for individual genes is limited. However, for well-powered genes (expected pLoF ≥ 10) (Supplementary Information), we observed a lower mean observed/expected ratio and LOEUF across genes among African/African American individuals, a population with a larger effective population size, compared with other populations (Extended Data Fig. 8d, e), consistent with the increased efficiency of selection in populations with larger effective population sizes[25,26].

Constraint informs disease aetiologies

The LOEUF metric can be applied to improve molecular diagnosis and advance our understanding of disease mechanisms. Disease-associated genes, discovered by different technologies over the course of many years across all categories of inheritance and effects, span the entire spectrum of LoF tolerance (Extended Data Fig. 9a). However, in recent years, high-throughput sequencing technologies have enabled the identification of highly deleterious variants that are de novo or only inherited in small families or trios, leading to the discovery of novel disease genes under extreme constraint against pLoF variation that could not have been identified by linkage approaches that rely on broadly inherited variation (Extended Data Fig. 9b). This result is consistent with a recent analysis that shows a post-whole-exome/whole-genome sequencing era enrichment for gene–disease relationships attributable to de novo variants[27].

Rare variants, which are more likely to be deleterious, are expected to exhibit stronger effects on average in constrained genes (previously shown using pLI from ExAC[28]), with an effect size related to the severity and reproductive fitness of the phenotype. In an independent cohort of 5,305 individuals with intellectual disability or developmental disorders and 2,179 controls, the rate of pLoF de novo variation in cases is 15-fold higher in genes belonging to the most constrained LOEUF decile, compared with controls (Fig. 5a), with a slightly increased rate (2.9-fold) in the second highest decile but not in others. A similar, but attenuated enrichment (4.4-fold in the most constrained decile) is seen for de novo variants in 6,430 patients with autism spectrum disorder (Extended Data Fig. 9c). Furthermore, in burden tests of rare variants (allele count across both cases and controls = 1) of patients with schizophrenia[28], we find a significantly higher odds ratio in constrained genes (Extended Data Fig. 9d).

Finally, although pLoF variants are predominantly rare, other more common variation in constrained genes may also be deleterious, including the effects of other coding or regulatory variants. In a heritability partitioning analysis of association results for 658 traits in the UK Biobank and other large-scale genome-wide association study (GWAS) efforts, we find an enrichment of common variant associations near genes that is linearly related to LOEUF decile across numerous traits (Fig. 5b). Schizophrenia and educational attainment are the most enriched traits (Fig. 5c), consistent with previous observations in associations between rare pLoF variants and these phenotypes[29–31]. This enrichment persists even when accounting for gene size, expression in GTEx brain samples, and previously tested annotations of functional

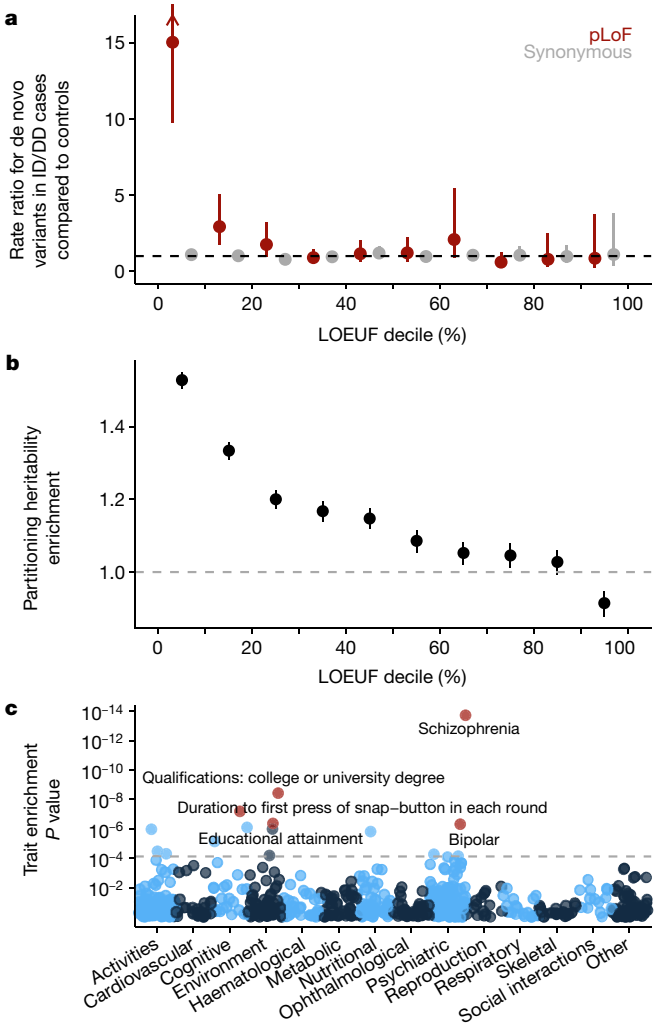


Fig. 5 | Disease applications of constraint. a, The rate ratio is defined by the rate of de novo variants (number per patient) in 5,305 cases of intellectual disability/developmental delay (ID/DD) divided by the rate in 2,179 controls. pLoF variants in the most constrained decile of the genome are approximately 11-fold more likely to be found in cases compared to controls. Error bars represent 95% confidence intervals. **b**, Marginal enrichment in per-SNV heritability explained by common (minor allele frequency > 5%) variants within 100-kb of genes in each LOEUF decile, estimated by linkage disequilibrium (LD) score regression[48]. Enrichment is compared to the average SNV genome-wide. The results reported here are from random effects meta-analysis of 276 independent traits (subsetted from the 658 traits with UK Biobank or large-scale consortium GWAS results). Error bars represent 95% confidence intervals. **c**, Conditional enrichment in per-SNV common variant heritability tested using regression of linkage disequilibrium score in each of 658 common disease and trait GWAS results. *P* values evaluate whether per-SNV heritability is proportional to the LOEUF of the nearest gene, conditional on 75 existing functional, linkage disequilibrium, and minor-allele-frequency-related genomic annotations. Colours alternate by broad phenotype category.

regions and evolutionary conservation, and suggests that some heritable polygenic diseases and traits, particularly cognitive or psychiatric ones, have an underlying genetic architecture that is driven substantially by constrained genes (Extended Data Fig. 10).

Discussion

In this paper and accompanying publications, we present the largest, to our knowledge, catalogue of harmonized variant data from any species so far, incorporating exome or genome sequence data from more than 140,000 humans. The gnomAD dataset of over 270 million variants is publicly available (https://gnomad.broadinstitute.org), and has already been widely used as a resource for estimates of allele frequency in the context of rare disease diagnosis (for a recent review, see Eilbeck et al.[32]), improving power for disease gene discovery[33–35], estimating genetic disease frequencies[36,37], and exploring the biological effect of genetic variation[38,39]. Here, we describe the application of this dataset to calculate a continuous metric that describes a spectrum of tolerance to pLoF variation for each protein-coding gene in the human genome. We validate this method using known gene sets and data from model organisms, and explore the value of this metric for investigating human gene function and discovery of disease genes.

We have focused on high-confidence, high-impact pLoF variants, calibrating our analysis to be highly specific to compensate for the increased false-positive rate among deleterious variants. However, some additional error modes may still exist, and indeed, several recent experiments have proposed uncharacterized mechanisms for escape from nonsense-mediated mRNA decay[40,41]. Furthermore, such a stringent approach will remove some true positives. For example, terminal truncations that are removed by LOFTEE may still exert a LoF mechanism through the removal of crucial C-terminal domains, despite the escape of the gene from nonsense-mediated decay. In addition, current annotation tools are incapable of detecting all classes of LoF variation and typically miss, for instance, missense variants that inactivate specific gene functions, as well as high-impact variants in regulatory regions. Future work will benefit from the increasing availability of high-throughput experimental assays that can assess the functional effect of all possible coding variants in a target gene[42], although scaling these experimental assays to all protein-coding genes represents a huge challenge. Identifying constraint in individual regulatory elements outside coding regions will be even more challenging, and require much larger sample sizes of whole genomes as well as improved functional annotation[43]. We discuss one class of high-impact regulatory variants in a companion manuscript[17], but many remain to be fully characterized.

Although the gnomAD dataset is of unprecedented scale, it has important limitations. At this sample size, we remain far from saturating all possible pLoF variants in the human exome; even at the most mutable sites in the genome (methylated CpG dinucleotides), we observe only half of all possible stop-gained variants. A substantial fraction of the remaining variants are likely to be heterozygous lethal, whereas others will exhibit an intermediate selection coefficient; much larger sample sizes (in the millions to hundreds of millions of individuals) will be required for comprehensive characterization of selection against all individual LoF variants in the human genome. Such future studies would also benefit substantially from increased ancestral diversity beyond the European-centric sampling of many current studies, which would provide opportunities to observe very rare and population-specific variation, as well as increase power to explore population differences in gene constraint. In particular, current reference databases including gnomAD have a near-complete absence of representation from the Middle East, central and southeast Asia, Oceania, and the vast majority of the African continent[44], and these gaps must be addressed if we are to fully understand the distribution and effect of human genetic variation.

It is also important to understand the practical and evolutionary interpretation of pLoF constraint. In particular, it should be noted that these metrics primarily identify genes undergoing selection against heterozygous variation, rather than strong constraint against homozygous variation[45]. In addition, the power of the LOEUF metric is affected by gene length, with approximately 30% of the coding genes in the genome still insufficiently powered for detection of constraint even at the scale of gnomAD (Fig. 2d). Substantially larger sample sizes and careful analysis of individuals enriched for homozygous pLoFs (see below) will be useful for distinguishing these possibilities. Furthermore, selection is largely blind to phenotypes emerging after reproductive age, and thus genes with phenotypes that manifest later in life, even if

severe or fatal, may exhibit much weaker intolerance to inactivation. Despite these caveats, our results demonstrate that pLoF constraint divides protein-coding genes in a way that correlates usefully with their probability of disease impact and other biological properties, and confirm the value of constraint in prioritizing candidate genes in studies of both rare and common diseases.

Examples such as *PCSK9* demonstrate the value of human pLoF variants for identifying and validating targets for therapeutic intervention across a wide range of human diseases. As discussed in more detail in an accompanying manuscript[12], careful attention must be paid to a variety of complicating factors when using pLoF constraint to assess candidates. More valuable information comes from directly exploring the phenotypic effect of LoF variants on carrier humans, both through 'forward genetics' approaches such as gene mapping to identify genes that cause Mendelian disease, as well as 'reverse genetics' approaches that leverage large collections of sequenced humans to find and clinically characterize individuals with disruptive mutations in specific genes. Although clinical data are currently available for only a small subset of gnomAD individuals, future efforts that integrate sequencing and deep phenotyping of large biobanks will provide valuable insight into the biological implications of partial disruption of specific genes. This is illustrated in a companion manuscript that explores the clinical correlates of heterozygous pLoF variants in the *LRRK2* gene, demonstrating that life-long partial inactivation of this gene is likely to be safe in humans[13].

Such examples, and the sheer scale of pLoF discovery in this dataset, suggest the near-future feasibility and considerable value of a human 'knockout' project—a systematic attempt to discover the phenotypic consequences of functionally disruptive mutations, in either the heterozygous or homozygous state, for all human protein-coding genes. Such an approach will require cohorts of samples from millions of sequenced and deeply, consistently phenotyped individuals and, for the discovery of 'complete' knockouts, would benefit substantially from the targeted inclusion of large numbers of samples from populations that have either experienced strong demographic bottlenecks or high levels of recent parental relatedness (consanguinity)[12]. Such a resource would allow the construction of a comprehensive map that directly links gene-disrupting variation to human biology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2308-7.

1. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
2. Schneeberger, K. Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat. Rev. Genet.* **15**, 662–676 (2014).
3. Zambrowicz, B. P. & Sands, A. T. Knockouts model the 100 best-selling drugs—will they model the next 100? *Nat. Rev. Drug Discov.* **2**, 38–51 (2003).
4. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
5. Chong, J. X. et al. The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).
6. Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr & Hobbs, H. H. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
7. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
8. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
9. Cassa, C. A. et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).
10. Petrovski, S. et al. The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet.* **11**, e1005492 (2015).
11. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* https://doi.org/10.1038/s41586-020-2287-8 (2020).
12. Minikel, E. V. et al. Evaluating drug targets through human loss-of-function genetic variation. *Nature* https://doi.org/10.1038/s41586-020-2267-z (2020).
13. Whiffin, N. et al. The effect of LRRK2 loss-of-function variants in humans. *Nature Med.* https://doi.org/10.1038/s41591-020-0893-5 (2020).
14. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
15. Cummings, B. B. et al. Transcript expression-aware annotation improves rare variant interpretation. *Nature* https://doi.org/10.1038/s41586-020-2329-2 (2020).
16. Wang, Q. et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat. Commun.* https://doi.org/10.1038/s41467-019-12438-5 (2020).
17. Whiffin, N. et al. Characterising the loss-of-function impact of 5′ untranslated region variants in whole genome sequence data from 15,708 individuals. *Nat. Commun.* https://doi.org/10.1038/s41467-019-10717-9 (2019).
18. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.19.33 (2013).
19. Hail Team. Hail 0.2.19; https://github.com/hail-is/hail/releases/tag/0.2.19 (released 2 August 2019).
20. Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
21. Motenko, H., Neuhauser, S. B., O'Keefe, M. & Richardson, J. E. MouseMine: a new data warehouse for MGI. *Mamm. Genome* **26**, 325–330 (2015).
22. Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A. & Richardson, J. E. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res.* **43**, D726–D736 (2015).
23. Hart, T. et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 (Bethesda)* **7**, 2719–2727 (2017).
24. Feiglin, A., Allen, B. K., Kohane, I. S. & Kong, S. W. Comprehensive analysis of tissue-wide gene expression and phenotype data reveals tissues affected in rare genetic disorders. *Cell Syst.* **5**, 140–148.e2 (2017).
25. Gravel, S. When is selection effective? *Genetics* **203**, 451–462 (2016).
26. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating the mutation load in human genomes. *Nat. Rev. Genet.* **16**, 333–343 (2015).
27. Bamshad, M. J., Nickerson, D. A. & Chong, J. X. mendelian gene discovery: fast and furious with no end in sight. *Am. J. Hum. Genet.* **105**, 448–455 (2019).
28. Walters, J. T. R. et al. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat. Genet.* **511**, 421 (2017).
29. Ganna, A. et al. Quantifying the impact of rare and uTheltra-rare coding variation across the phenotypic spectrum. *Am. J. Hum. Genet.* **102**, 1204–1211 (2018).
30. Ganna, A. et al. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.* **19**, 1563–1565 (2016).
31. Genovese, G. et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
32. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).
33. DeBoever, C. et al. Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).
34. Emdin, C. A. et al. Analysis of predicted loss-of-function variants in UK Biobank identifies variants protective for disease. *Nat. Commun.* **9**, 1613 (2018).
35. Satterstrom, F. K. et al. Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants. *Nat. Neurosci.* **22**, 1961–1965 (2019).
36. de Andrade, K. C. et al. Variable population prevalence estimates of germline TP53 variants: a gnomAD-based analysis. *Hum. Mutat.* **40**, 97–105 (2019).
37. Laver, T. W. et al. Analysis of large-scale sequencing cohorts does not support the role of variants in UCP2 as a cause of hyperinsulinaemic hypoglycaemia. *Hum. Mutat.* **38**, 1442–1444 (2017).
38. Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
39. Glassberg, E. C., Lan, X. & Pritchard, J. K. Evidence for weak selective constraint on human gene expression. *Genetics* **211**, 757–772 (2019).
40. El-Brolosy, M. A. et al. Genetic compensation triggered by mutant mRNA degradation. *Nature* **568**, 193–197 (2019).
41. Tuladhar, R. et al. CRISPR-Cas9-based mutagenesis frequently provokes on-target mRNA misregulation. *Nat. Commun.* **10**, 4056 (2019).
42. Findlay, G. M. et al. Accurate classification of *BRCA1* variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
43. Short, P. J. et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).
44. Martin, A. R., Kanai, M., Kamatani, Y., Neale, B. M. & Daly, M. J. Hidden 'risk' in polygenic scores: clinical use today could exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
45. Fuller, Z., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
46. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
47. Diaz-Papkovich, A., Anderson-Trocme, L. & Gravel, S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet.* (2018). https://doi.org/10.1371/journal.pgen.1008432
48. Finucane, H. K. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
49. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
50. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
51. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
52. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Naturey* **485**, 242–245 (2012).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Genome Aggregation Database Consortium

Carlos A. Aguilar Salinas[20], Tariq Ahmad[21], Christine M. Albert[22,23], Diego Ardissino[24], Gil Atzmon[25,26,27], John Barnard[28], Laurent Beaugerie[29], Emelia J. Benjamin[30,31,32], Michael Boehnke[33], Lori L. Bonnycastle[34], Erwin P. Bottinger[35], Donald W. Bowden[36,37,38], Matthew J. Bown[39,40], John C. Chambers[41,42,43], Juliana C. Chan[44], Daniel Chasman[22,45], Judy Cho[35], Mina K. Chung[28], Bruce Cohen[45,46], Adolfo Correa[47], Dana Dabelea[48], Mark J. Daly[1,2,9], Dawood Darbar[49], Ravindranath Duggirala[50], Josée Dupuis[30,51], Patrick T. Ellinor[1,52], Roberto Elosua[53,54,55], Jeanette Erdmann[56,57,58], Tõnu Esko[1,59], Martti Färkkilä[60], Jose Florez[1,7,61,62,63], Andre Franke[64], Gad Getz[45,65,66,67,68], Benjamin Glaser[69], Stephen J. Glatt[70], David Goldstein[71,72], Clicerio Gonzalez[73], Leif Groop[6,74], Christopher Haiman[75], Craig Hanis[76], Matthew Harms[77,78], Mikko Hiltunen[79], Matti M. Holi[80], Christina M. Hultman[81,82], Mikko Kallela[83], Jaakko Kaprio[6,84], Sekar Kathiresan[5,45,85], Bong-Jo Kim[86], Young Jin Kim[86], George Kirov[87], Jaspal Kooner[10,41,42,43], Seppo Koskinen[88], Harlan M. Krumholz[89], Subra Kugathasan[90], Soo Heon Kwak[91], Markku Laakso[92,93], Terho Lehtimäki[94], Ruth J. F. Loos[35,95], Steven A. Lubitz[1,52], Ronald C. W. Ma[44,96,97], Daniel G. MacArthur[1,2], Jaume Marrugat[54,98], Kari M. Mattila[94], Steven McCarroll[9,99], Mark I. McCarthy[100,101,102], Dermot McGovern[103], Ruth McPherson[104], James B. Meigs[1,45,105], Olle Melander[106], Andres Metspalu[59], Benjamin M. Neale[1,2], Peter M. Nilsson[107], Michael C. O'Donovan[87], Dost Ongur[45,46], Lorena Orozco[108], Michael J. Owen[87], Colin N. A. Palmer[109], Aarno Palotie[1,6,9], Kyong Soo Park[91,110], Carlos Pato[111], Ann E. Pulver[112], Nazneen Rahman[113], Anne M. Remes[114], John D. Rioux[115,116], Samuli Ripatti[1,6,84], Dan M. Roden[117,118], Danish Saleheen[119,120,121], Veikko Salomaa[122], Nilesh J. Samani[39,40], Jeremiah Scharf[1,5,9], Heribert Schunkert[123,124], Moore B. Shoemaker[125], Pamela Sklar[62,126,127,128,151], Hilkka Soininen[129], Harry Sokol[29], Tim Spector[130], Patrick F. Sullivan[81,131], Jaana Suvisaari[122], E. Shyong Tai[132,133,134], Yik Ying Teo[132,135,136], Tuomi Tiinamaija[6,137,138], Ming Tsuang[139,140], Dan Turner[141], Teresa Tusie-Luna[142,143], Erkki Vartiainen[84], James S. Ware[1,10,11], Hugh Watkins[144], Rinse K. Weersma[145], Maija Wessman[6,137], James G. Wilson[146] & Ramnik J. Xavier[147,148]

[20]Unidad de Investigacion de Enfermedades Metabolicas, Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico. [21]Peninsula College of Medicine and Dentistry, Exeter, UK. [22]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA. [23]Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [24]Department of Cardiology, University Hospital, Parma, Italy. [25]Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel. [26]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, USA. [27]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. [28]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [29]Sorbonne Université, APHP, Gastroenterology Department, Saint Antoine Hospital, Paris, France. [30]Framingham Heart Study, National Heart, Lung, & Blood Institute and Boston University, Framingham, MA, USA. [31]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [32]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA. [33]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [34]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [35]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [36]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [37]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [38]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [39]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK. [40]NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK. [41]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [42]Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK. [43]Imperial College Healthcare NHS Trust, Imperial College London, London, UK. [44]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. [45]Department of Medicine, Harvard Medical School, Boston, MA, USA. [46]Program for Neuropsychiatric Research, McLean Hospital, Belmont, MA, USA. [47]Department of Medicine, University of Mississippi Medical Center, Jackson, MI, USA. [48]Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA. [49]Department of Medicine and Pharmacology, University of Illinois at Chicago, Chicago, IL, USA. [50]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA. [51]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [52]Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [53]Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain. [54]Centro de Investigación Biomédica en Red Enfermedades Cardiovaculares (CIBERCV), Barcelona, Catalonia, Spain. [55]Department of Medicine, Medical School, University of Vic-Central University of Catalonia, Vic, Catalonia, Spain. [56]Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany. [57]DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, Lübeck, Germany. [58]University Heart Center Lübeck, Lübeck, Germany. [59]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [60]Helsinki University and Helsinki University Hospital, Clinic of Gastroenterology, Helsinki, Finland. [61]Diabetes Unit, Massachusetts General Hospital, Boston, MA, USA. [62]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [63]Program in Metabolism, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [64]Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany. [65]Bioinformatics Consortium, Massachusetts General Hospital, Boston, MA, USA. [66]Cancer Genome Computational Analysis Group, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [67]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [68]Cancer Center, Massachusetts General Hospital, Boston, MA, USA. [69]Endocrinology and Metabolism Department, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [70]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA. [71]Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [72]Department of Genetics and Development, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [73]Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Cuernavaca, Mexico. [74]Genomics, Diabetes and Endocrinology, Lund University, Lund, Sweden. [75]Lund University Diabetes Centre, Malmö, Sweden. [76]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA. [77]Department of Neurology, Columbia University, New York, NY, USA. [78]Institute of Genomic Medicine, Columbia University, New York, NY, USA. [79]Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland. [80]Department of Psychiatry, Helsinki University Central Hospital, Lapinlahdentie, Helsinki, Finland. [81]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [82]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [83]Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland. [84]Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [85]Cardiovascular Disease Initiative and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [86]Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, South Korea. [87]MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Cardiff, UK. [88]Department of Health, National Institute for Health and Welfare (THL), Helsinki, Finland. [89]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA. [90]Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, GA, USA. [91]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. [92]Institute of Clinical Medicine, The University of Eastern Finland, Kuopio, Finland. [93]Kuopio University Hospital, Kuopio, Finland. [94]Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [95]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [96]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [97]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. [98]Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain. [99]Department of Genetics, Harvard Medical School, Boston, MA, USA. [100]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Headington, Oxford, UK. [101]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [102]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [103]F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [104]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada. [105]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [106]Department of Clinical Sciences, University Hospital Malmo Clinical Research Center, Lund University, Malmo, Sweden. [107]Department of Clinical Sciences, Lund University, Skane University Hospital, Malmo, Sweden. [108]Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, Mexico. [109]Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. [110]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea. [111]Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, CA, USA. [112]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [113]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK. [114]Medical Research Center, Oulu University Hospital, Oulu, Finland and Research Unit of Clinical Neuroscience, Neurology, University of Oulu, Oulu, Finland. [115]Research Center, Montreal Heart Institute, Montreal, Quebec, Canada. [116]Department of Medicine, Faculty of Medicine, Université de Montréal, Quebec, Canada. [117]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [118]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [119]Department of Biostatistics and

Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [120]Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [121]Center for Non-Communicable Diseases, Karachi, Pakistan. [122]National Institute for Health and Welfare, Helsinki, Finland. [123]Deutsches Herzzentrum München, Munich, Germany. [124]Technische Universität München, Munich, Germany. [125]Division of Cardiovascular Medicine, Nashville VA Medical Center and Vanderbilt University, School of Medicine, Nashville, TN, USA. [126]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [127]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [128]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [129]Institute of Clinical Medicine, Neurology, University of Eastern Finlad, Kuopio, Finland. [130]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [131]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [132]Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore, Singapore. [133]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [134]Duke-NUS Graduate Medical School, Singapore, Singapore. [135]Life Sciences Institute, National University of Singapore, Singapore, Singapore. [136]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore. [137]Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland. [138]HUCH Abdominal Center, Helsinki University Hospital, Helsinki, Finland. [139]Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego, CA, USA. [140]Institute of Genomic Medicine, University of California, San Diego, CA, USA. [141]Juliet Keidan Instit.ute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel. [142]Instituto de Investigaciones Biomédicas UNAM, Mexico City, Mexico. [143]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. [144]Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [145]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. [146]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [147]Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [148]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. [151]Deceased: Pamela Sklar.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The gnomAD 2.1.1 dataset is available for download at http://gnomad.broadinstitute.org, where we have developed a browser for the dataset and provide files with detailed frequency and annotation information for each variant. There are no restrictions on the aggregate data released.

Code availability

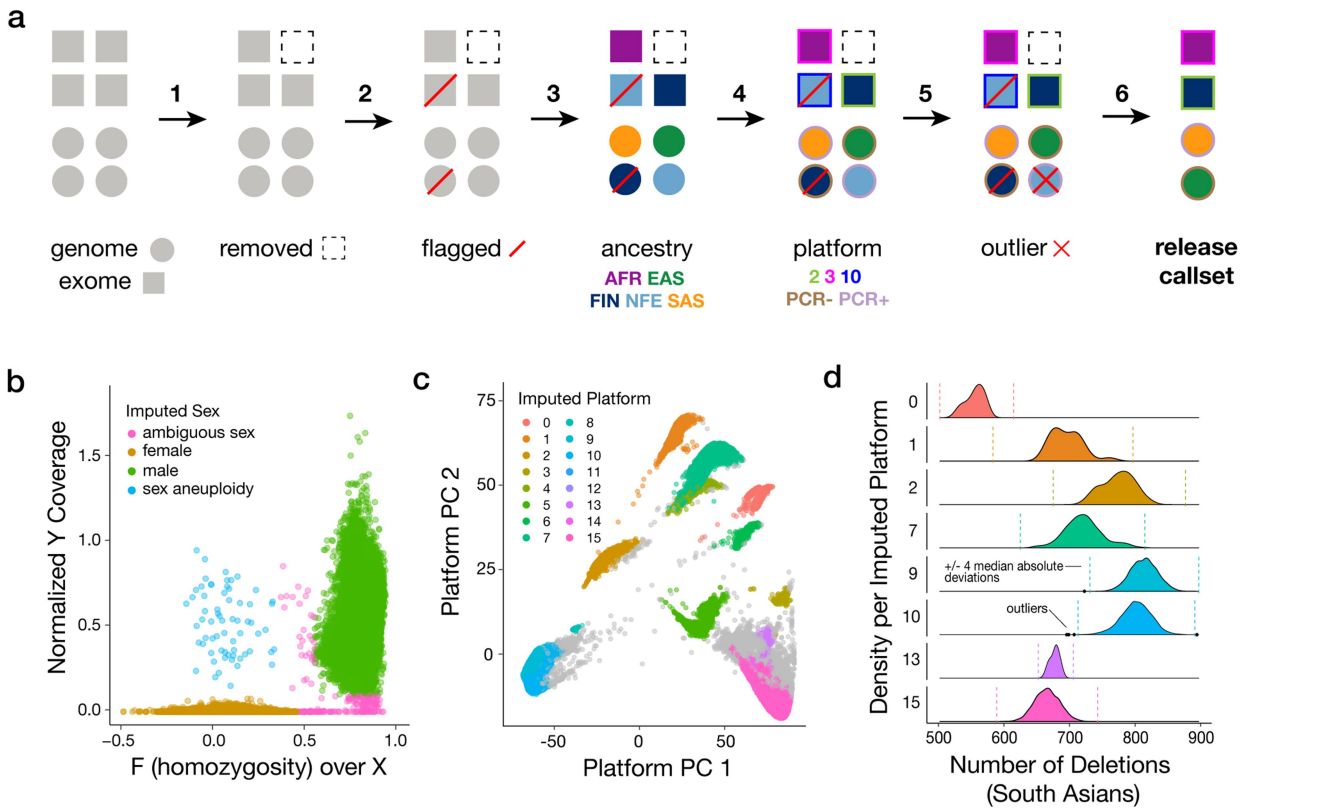
All code to perform quality control is provided at https://github.com/broadinstitute/gnomad_qc, and the code to perform all analyses and regenerate all the figures in this manuscript is provided at https://github.com/macarthur-lab/gnomad_lof. LOFTEE is available at https://github.com/konradjk/loftee. All code and software to reproduce figures are available in a Docker image at konradjk/gnomad_lof_paper:0.2.

Acknowledgements We thank the many individuals whose sequence data are aggregated in gnomAD for their contributions to research, and the users of gnomAD for their collaborative feedback. We also thank D. Altshuler for contributions to the development of the gnomAD resource, and A. Martin, E. Fauman, J. Bloom, D. King and the Hail team for discussions. The results published here are in part based on data: (1) generated by The Cancer Genome Atlas (TCGA) managed by the NCI and NHGRI (accession: phs000178.v10.p8); information about TCGA can be found at http://cancergenome.nih.gov; (2) generated by the Genotype-Tissue Expression Project (GTEx) managed by the NIH Common Fund and NHGRI (accession: phs000424.v7.p2); (3) generated by the Exome Sequencing Project, managed by NHLBI; (4) generated by the Alzheimer's Disease Sequencing Project (ADSP), managed by the NIA and NHGRI (accession: phs000572.v7.p4). K.J.K. was supported by NIGMS F32 GM115208. L.C.F. was supported by the Swiss National Science Foundation (Advanced Postdoc.Mobility 177853). J.X.C. was supported by NHGRI and NHLBI grants UM1 HG006493 and U24 HG008956. Analysis of the Genome Aggregation Database was funded by NIDDK U54 DK105566, NHGRI UM1 HG008900, BioMarin Pharmaceutical Inc., and Sanofi Genzyme Inc. Development of LOFTEE was funded by NIGMS R01 GM104371. D.G.M, K.M.L, and M.E.T. were supported by NICHD HD081256. D.G.M., R.L.C. and M.E.T. were supported by NIMH MH115957. The complete acknowledgments can be found in the Supplementary Information. We have complied with all relevant ethical regulations.

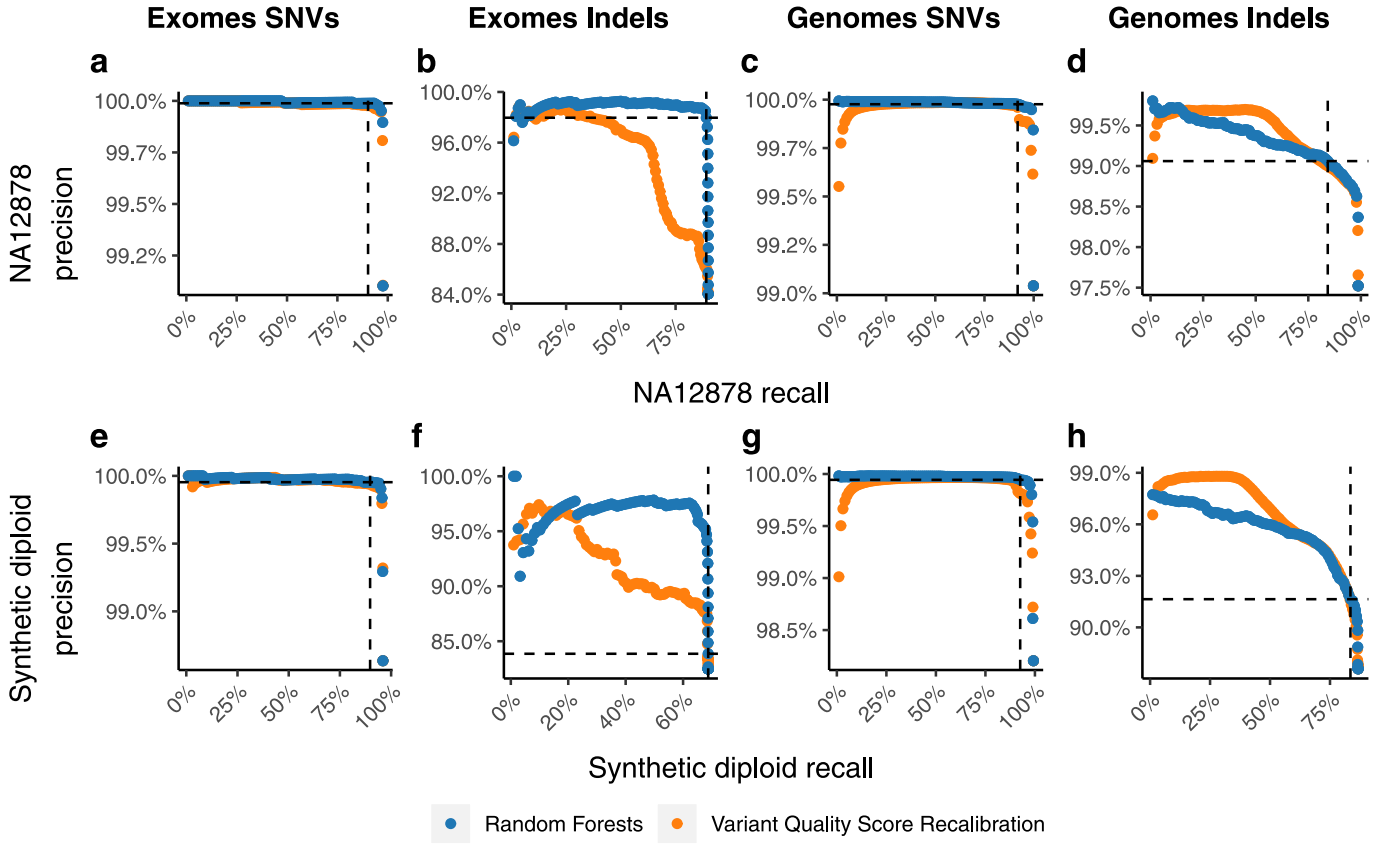
Author contributions K.J.K., L.C.F., G.T., B.B.C., J.A., Q.W., R.L.C., K.M.L., A.G., M.S., D.R., M.S.-B., B.M.N., M.J.D. and D.G.M. contributed to the writing of the manuscript and generation of figures. K.J.K., L.C.F., G.T., B.B.C., Q.W., R.L.C., K.M.L., A.G., H.B., D.R., M.S.-B., E.M.E., E.G.S., J.A.K., N.W., J.X.C., K.E.S., E.P.-H., Z.Z., A.H.O'D.-L., M.E.T., B.M.N., M.J.D. and D.G.M. contributed to the analysis of data. K.J.K., L.C.F., G.T., B.B.C., K.M.L., D.P.B., L.D.G., M.S., N.A.W., R.K.W., K.T., Y.F., E.B., T.P., A.W., C.S., K.E.S., Z.Z., A.H.O'D.-L., C.V., B.M.N., M.J.D. and D.G.M. developed tools and methods that enabled the scientific discoveries herein. K.J.K., L.C.F., G.T., B.B.C., J.A., R.L.C., K.M.L., L.D.G., Y.F., E.B., A.H.O'D.-L., E.V.M., B.W., M.L., J.S.W., C.V., I.M.A., L.B., K.C., K.M.C., M.C., S.D., S.F., S.G., J.G., N.G., T.J., D.K., C.L., R.M., S.N., N.P., D.R., V.R.-R., A.S., M.S., J.S., K.T., C.T., G.W., M.E.T., B.M.N., M.J.D. and D.G.M. contributed to the production and quality control of the gnomAD dataset. All authors listed under The Genome Aggregation Database Consortium contributed to the generation of the primary data incorporated into the gnomAD resource. All authors reviewed the manuscript.

Competing interests K.J.K. owns stock in Personalis. R.K.W. has received unrestricted research grants from Takeda Pharmaceutical Company. A.H.O'D.-L. has received honoraria from ARUP and Chan Zuckerberg Initiative. E.V.M. has received research support in the form of charitable contributions from Charles River Laboratories and Ionis Pharmaceuticals, and has consulted for Deerfield Management. J.S.W. is a consultant for MyoKardia. B.M.N. is a member of the scientific advisory board at Deep Genomics and consultant for Camp4 Therapeutics, Takeda Pharmaceutical, and Biogen. M.J.D. is a founder of Maze Therapeutics. D.G.M. is a founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme. The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. M.I.M. has served on advisory panels for Pfizer, NovoNordisk, Zoe Global; has received honoraria from Merck, Pfizer, NovoNordisk and Eli Lilly; has stock options in Zoe Global and has received research funding from Abbvie, Astra Zeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, Servier & Takeda. As of June 2019, M.I.M. is an employee of Genentech, and holds stock in Roche. N.R. is a non-executive director of AstraZeneca.

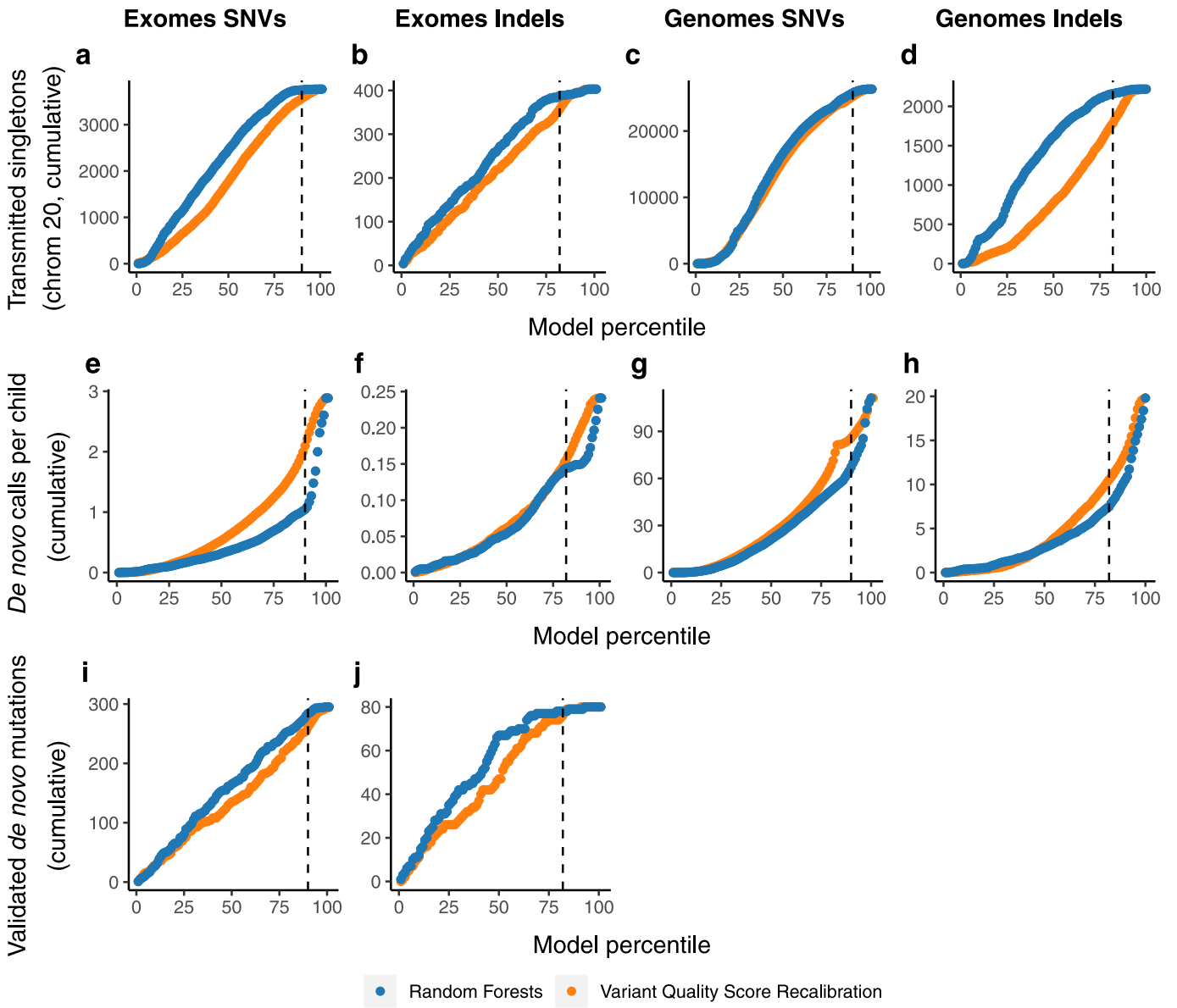
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2308-7.
Correspondence and requests for materials should be addressed to K.J.K. or D.G.M.
Peer review information *Nature* thanks Deanna Church, Rayna Harris, Alexander Hoischen and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.



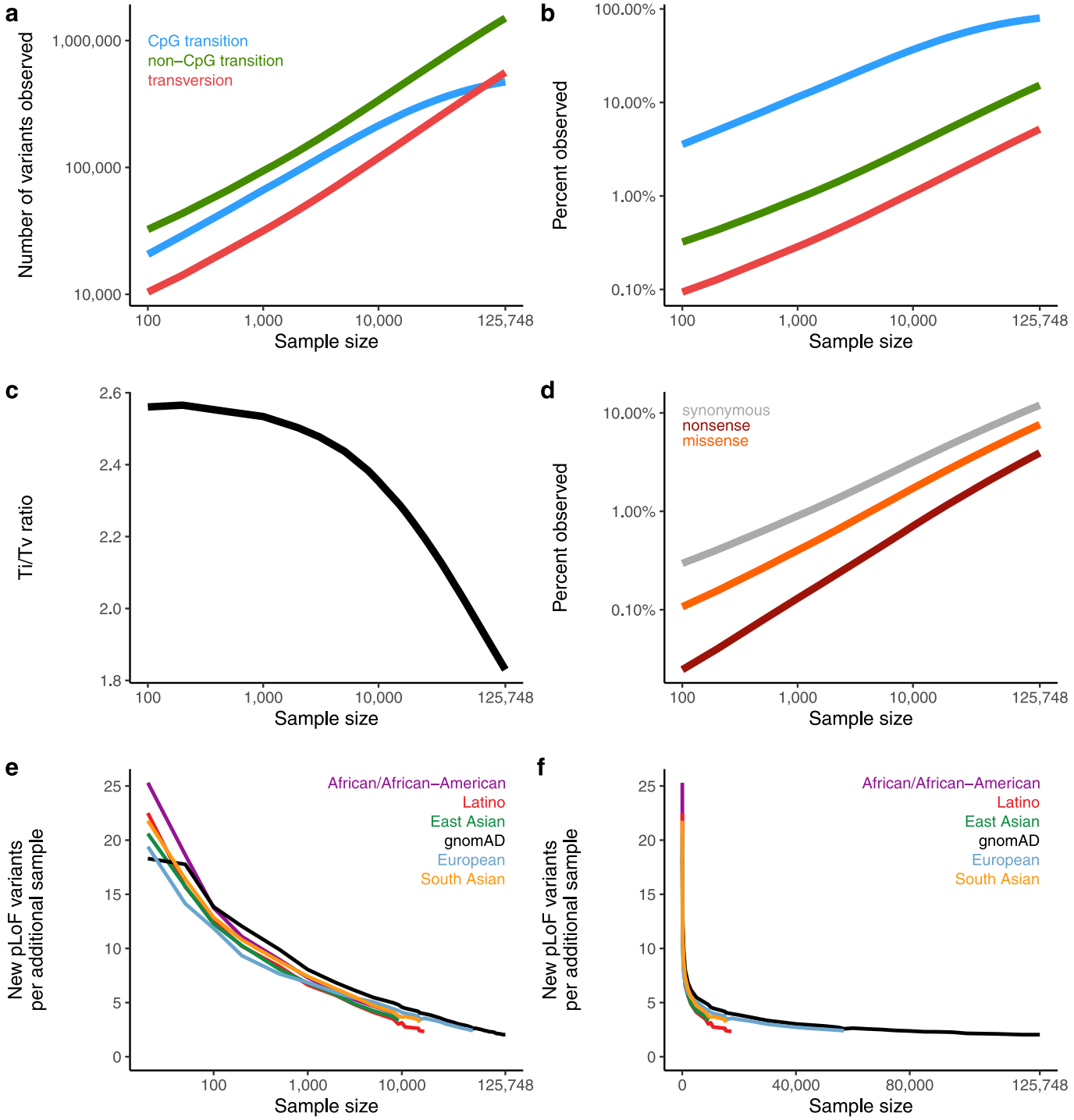
Extended Data Fig. 1 | Overview of the sample quality control workflow. **a**, Exome (square) and genome (circle) samples underwent quality control in the following stages: hard filtering (step 1), relatedness inference (step 2), ancestry inference (step 3), platform inference (step 4, for exomes only), and population- and platform-specific outlier filtering (step 5). See Supplementary Information for further details. Except for samples failing hard filters (dotted outline), all quality control analyses were applied to all samples, regardless of the presence or absence of other quality control flags (such as relatedness, lack of release permissions, or outlier status; red diagonal bar). Assignment of ancestry labels is represented by fill colour and accompanying three-letter ancestry group abbreviation. Assignment of platform labels is represented by outline colour and a numbered label for exomes (corresponding to imputed platforms) and a PCR ± label for genomes. The final set of samples included in the gnomAD release (125,748 exomes and 15,708 genomes) was defined to be the set of unrelated samples with release permissions, no hard filter flags, and no population- and platform-specific outlier metrics (step 6). **b**, In exomes, the chromosomal sex of samples was inferred based on the inbreeding coefficient on chromosome X and the coverage of chromosome Y into male (green), female (amber), ambiguous sex (pink), and sex chromosome aneuploid (blue). **c**, The top two principal components from PCA-HDBSCAN analysis of exome capture regions. Sequencing platforms were inferred for exome samples based on principal component analysis of biallelic variant call rates over all known exome capture regions, and samples were assigned a cluster label (0–15, or unknown) using HDBSCAN. **d**, We performed platform- and population-specific outlier filtering for several quality-control metrics. The distribution of the number of deletions in samples from south Asian individuals across platforms is shown. Distributions (and accordingly, median and median absolute deviations) for these metrics varied widely both by population and sequencing platform (numbered on the *y* axis). Outliers (black dots) were defined as samples with values outside four median absolute deviations (shown by dotted vertical lines) from the median of a given metric.



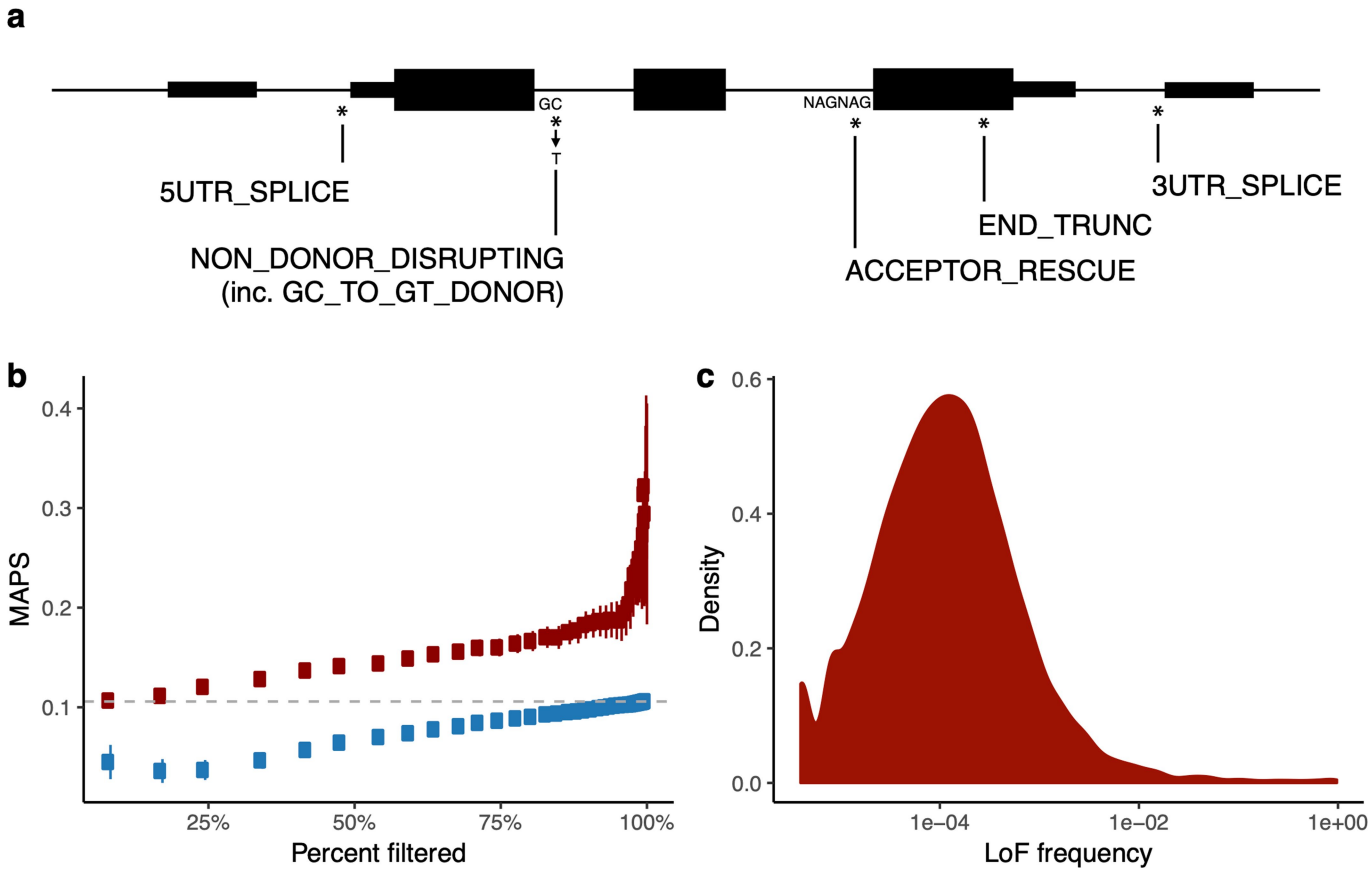
Extended Data Fig. 2 | Variant calling performance for common variants. **a**–**h**, Precision-recall curves are shown for variant calls in two samples with independent gold-standard data, NA12878[49] (**a**–**d**) and a synthetic diploid mixture[50] (**e**–**h**). The random forest (blue) approach described here is compared to the current state-of-the-art GATK variant quality score recalibration (orange) for exome SNVs (**a**, **e**) and indels (**b**, **f**), and genome SNVs (**c**, **g**) and indels (**d**, **h**). Note that the indels presented in **f** and **h** exclude 1-base-pair (bp) indels as they are not well characterized in the synthetic diploid mixture gold standard sample. In all cases, at the thresholds chosen (dashed lines representing 10% and 20% of SNVs and indels filtered, respectively), random forest outperforms or is similar to variant quality score recalibration.



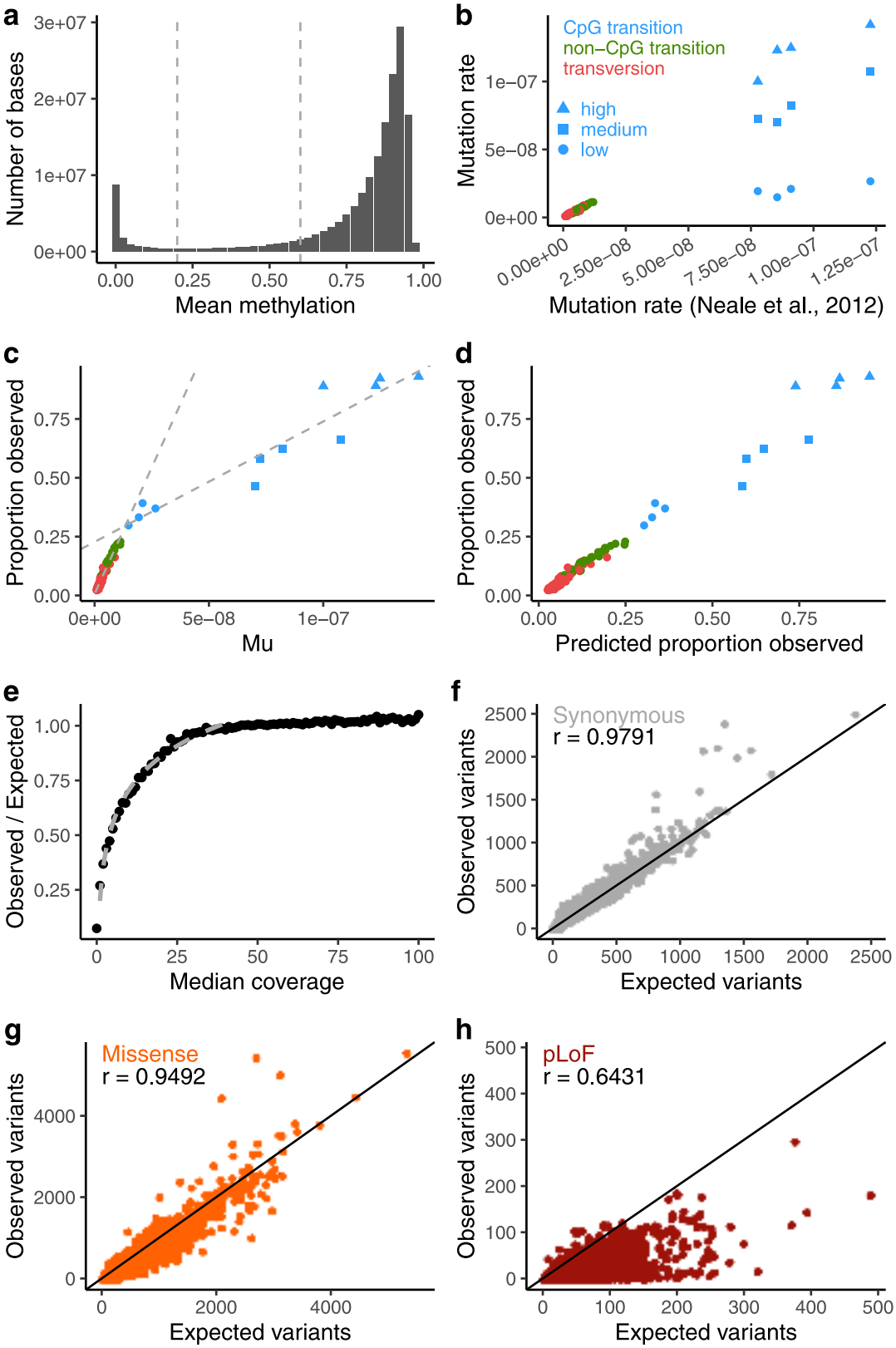
Extended Data Fig. 3 | Variant calling performance for rare variants. **a**–**j**, The *x* axes show the cumulative ranked percentile for our random forest (blue) model and, as a comparison, for the current state-of-the-art GATK variant quality score recalibration (orange). That is, the point at 10 shows the performance of the 10% best-scored data; the point at 50 shows the performance 50% best-scored data. **a**–**d**, The number of transmitted singletons (singletons in the unrelated individuals that are transmitted to an offspring) on chromosome 20 for exome SNVs (**a**) and indels (**b**), and genome SNVs (**c**) and indels (**d**). Chromosome 20 was not used for training our random forest model. We expect most of these to be real variants because we observe Mendelian transmission of an allele that was sequenced independently in a parent and child. **e**–**h**, The number of bi-allelic de novo calls per child (4,568 exomes, 212 genomes) outside of low-complexity regions. The expectation is that there is approximately 1.6 de novo SNV (**e**) and 0.1 de novo indels per exome (**f**), and 65 de novo SNVs (**g**) and 5 de novo indels (**h**) per genome[20]. **i**, **j**, The number of independently validated de novo mutations, available for a subset of 331 exome samples for which de novo mutations were validated as part of other studies[51]. In all cases, at the thresholds chosen (dashed lines representing 10% and 20% of SNVs and indels filtered, respectively), random forest outperforms or is similar to variant quality score recalibration.



Extended Data Fig. 4 | Variant discovery at large sample sizes. a, **b**, The total number of variants observed (**a**) and the proportion of possible variants observed (**b**) as a function of sample size, broken down by variant class. At large sample sizes, CpG transitions become saturated, as previously described[4]. Colours are consistent in **a** and **b**. **c**, This results in a decrease of the transition/transversion (Ti/Tv) ratio. **d**, When broken down by functional class, we observe the effects of selection, in which synonymous variants have the highest proportion observed, followed by missense and pLoF variants. **e**, **f**, The number of additional pLoF variants introduced into the cohort as a function of sample size on a log (**e**) and linear (**f**) scale. Here, gnomAD (black) refers to a uniform sampling from the population distribution of the full cohort of exome-sequenced individuals.

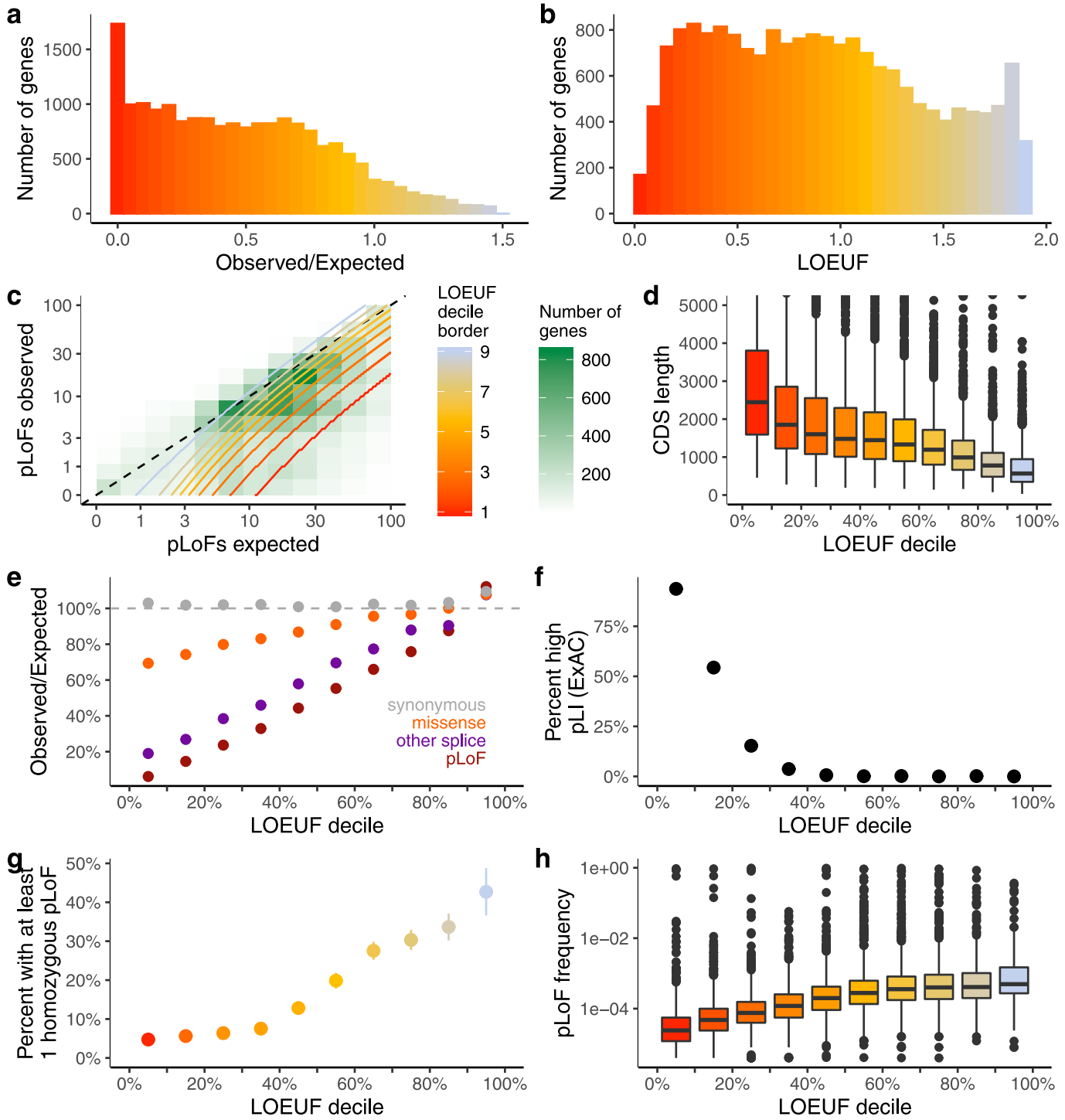


Extended Data Fig. 5 | Using LOFTEE to create a high-confidence set of pLoF variation. a, Schematic of LOFTEE filters. LOFTEE filters out putative stop-gained, essential splice, and frameshift variants based on sequence and transcript context, as well as flagging exonic features such as conservation (not shown). For instance, variants that are not predicted to disrupt splicing based on retention of a strong splice site, or rescue of a nearby splice site. Additional filters not shown include: ANC_ALLELE (the alternative allele is the ancestral allele), NON_ACCEPTOR_DISRUPTING and DONOR_RESCUE (opposite to those already shown). **b**, To tune the END_TRUNC filter, we retained variants that pass the 50-bp rule (are more than 50 bp before the 3′-most splice site). The overall MAPS score for variants that fail this rule is shown in grey. For the remaining 39,072 variants, we computed the sum of the genomic evolutionary rate profiling (GERP) score of bases deleted by the variant. At 40 bins of this score, we compute the MAPS score for those variants retained at this threshold (red) compared to variants removed at this threshold (blue), and plot this as a function of the proportion of variants filtered at this threshold. We chose the 50% point as it retains variants with a MAPS score of 0.14, while removing variants with a MAPS score of 0.06. Error bars represent 95% confidence intervals. **c**, Density plot of aggregate pLoF frequency computed from high-confidence pLoF variants discovered using LOFTEE.

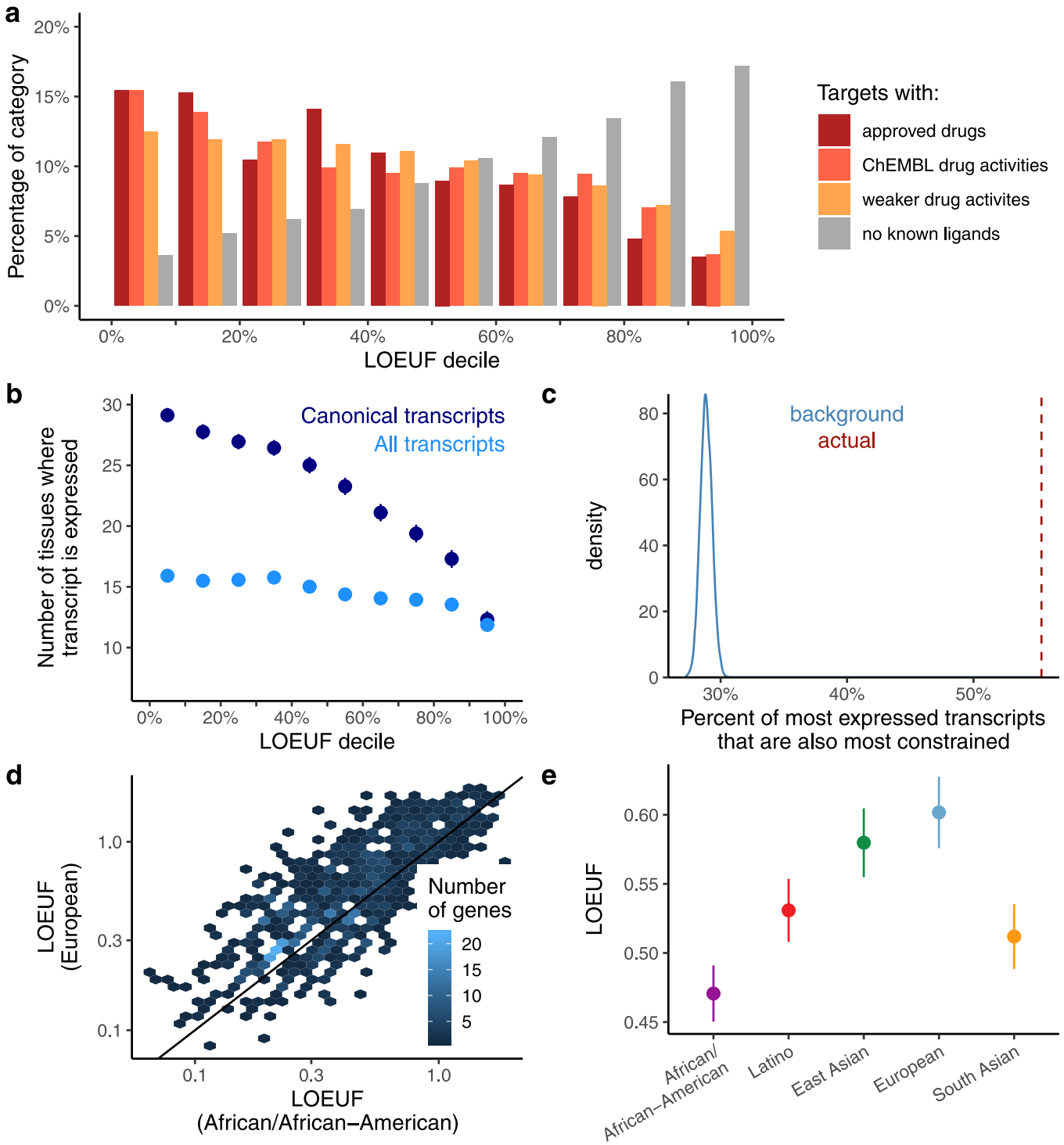


Extended Data Fig. 6 | See next page for caption.

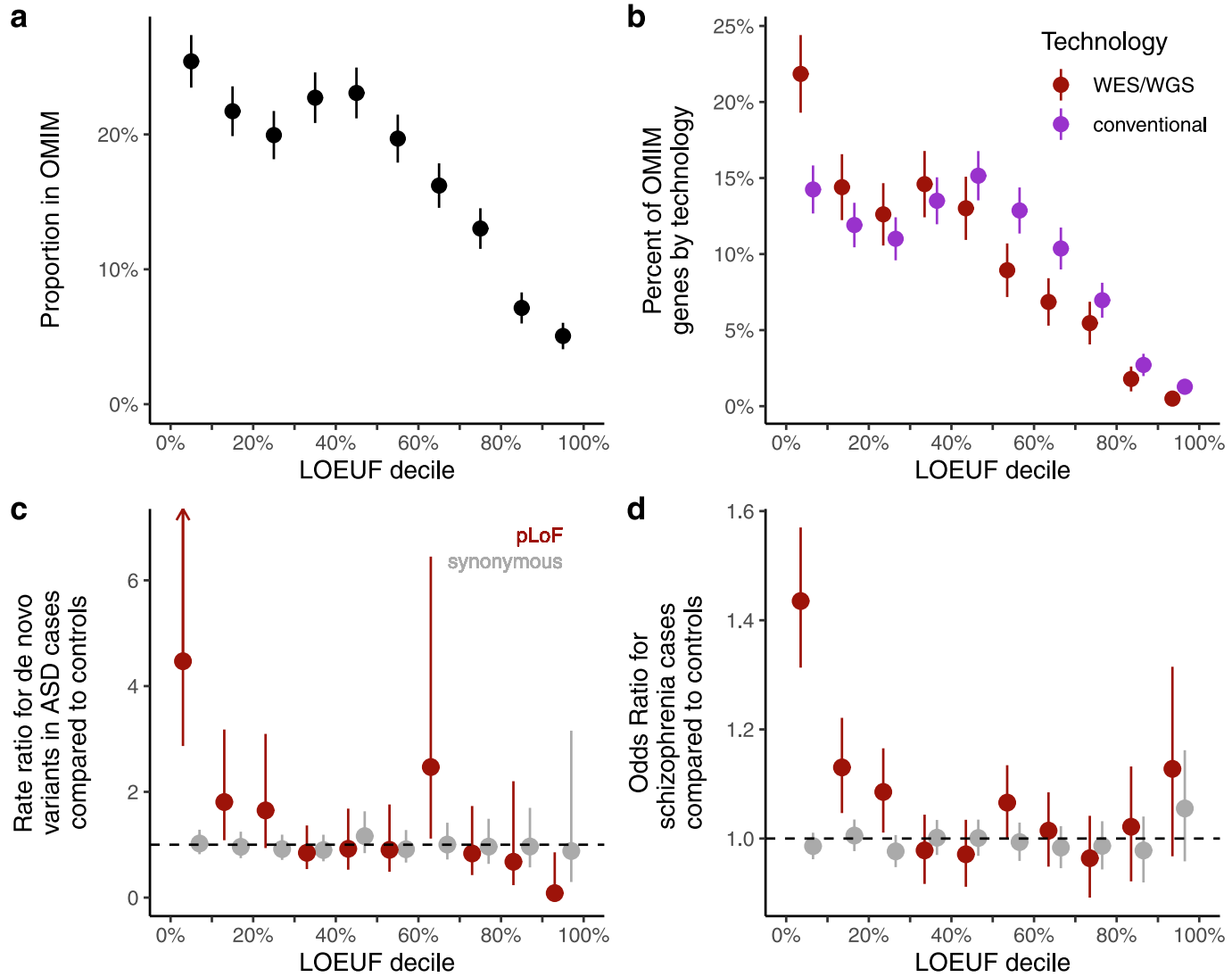
Extended Data Fig. 6 | Computing the depletion of variation of functional categories. a, The distribution of mean methylation values across 37 tissues and across every CpG dinucleotide in the genome. We divided the genome into 3 levels (low methylation, missing or < 0.2; medium, 0.2–0.6; and high, >0.6) and computed all ensuing metrics based on these categories. **b**, Comparison of estimates of the mutation rate with previous estimates[52]. For transversions and non-CpG transitions, we observe a strong correlation (linear regression $r = 0.98$; $P = 2.6 \times 10^{-65}$). For CpG transitions, the new estimates are calculated separately for the three levels of methylation and track with these levels. Colours and shapes are consistent in **b**–**d**. **c**, For **c**–**e**, only synonymous variants are considered. The proportion of possible variants observed for each context is correlated with the mutation rate. We compute two fit lines, one for CpG transitions, and one for other contexts to calibrate our estimates. **d**, Calibration of each context to compute a predicted proportion observed after fitting the two models in **c**, which is used to calculate an expected number of variants at high coverage. **e**, With an expectation computed from high coverage regions, the observed/expected ratio follows a logarithmic trend with the median coverage below 40×, which is used to correct low coverage bases in the final expectation model. **f**–**h**, For each transcript, the observed number of variants is plotted against the expected number from the model described above, for synonymous (**f**), missense (**g**), and pLoF (**h**) variants, and the linear regression coefficient is shown. Note that the expectation does not include selection, and so, pLoF and, to a lesser extent, missense variants exhibit lower observed values than expected.



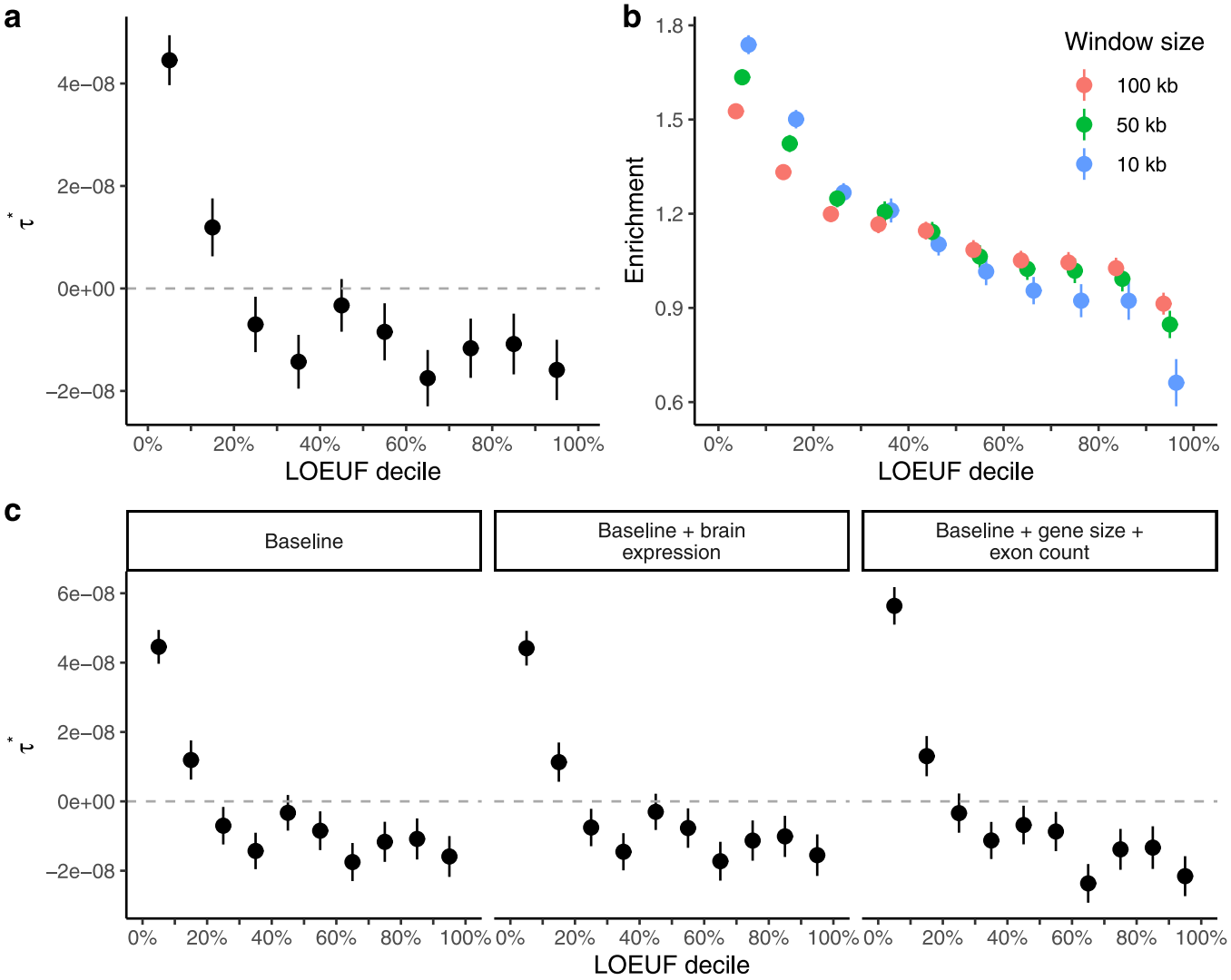
Extended Data Fig. 7 | Genomic properties of constrained genes. **a**, **b**, Histogram of the observed/expected ratio of pLoF variation (**a**) and LOEUF (**b**). Most genes have fewer observed variants than expected (median observed/expected = 0.48), and the genes with no observed pLoFs are distinguished between confidently constrained genes and noise by LOEUF. **c**, A 2D density plot of the number of observed versus expected pLoF variants. The boundaries of each decile are plotted as gradients (that is, the most constrained decile is below the lowest red line). **d**, The LOEUF of a gene is correlated with its coding sequence length (beta = -1.07×10^{-4}; $P < 10^{-100}$): thus, for all downstream statistical tests, we adjust for gene length or remove genes with fewer than 10 expected pLoFs. **e**, Observed/expected ratios of various functional classes across genes within each LOEUF decile. The most constrained decile has approximately 6% of the expected pLoFs, while synonymous variants are not depleted and missense variants exhibit modest depletion. **f**, The percentage of each LOEUF decile that was described in ExAC as constrained, or pLI > 0.9[4]. **g**, The percentage of each LOEUF decile that have at least one homozygous pLoF variant. **h**, Box plots of the aggregate pLoF frequency for each LOEUF decile. Centre line denotes the median; box limits denote upper and lower quartiles; whiskers denote 1.5× the interquartile range; points denote outliers). In **e**–**g**, error bars represent 95% confidence intervals (note that in some cases these are fully contained within the plotted point).



Extended Data Fig. 8 | Biological properties of constrained genes. a, The percentage of genes in each functional category from Pharos (see Supplementary Information) is broken down by the LOEUF decile. **b**, The mean number of tissues in which a transcript is expressed, binned by transcript-based LOEUF decile, is shown for all transcripts and canonical transcripts. **c**, The percentage of genes in which the most expressed transcript is also the most constrained is plotted in red, which is enriched compared to a permuted set (blue). **d**, For 927 genes with expected pLoF ≥10 in both the African/African American and European population subsets ($n = 8{,}128$), the LOEUF scores are highly correlated (linear regression $r = 0.78$, $P < 10^{-100}$), with a lower mean score observed in the African/African American population (0.49 versus 0.62; two-sided t-test $P = 4.1 \times 10^{-14}$), which has a higher effective population size. **e**, The mean LOEUF score for 865 genes with expected pLoF ≥10 in all populations ($n = 8{,}128$). Error bars represent 95% confidence intervals.



Extended Data Fig. 9 | Applications of constraint metrics to rare variant analysis of disease. a, Proportion of each LOEUF decile found in OMIM. **b**, Proportion of disease-associated genes discovered by whole-exome/genome sequencing (WES/WGS) compared to conventional (typically linkage) methods, plotted by LOEUF decile. The former are more constrained (LOEUF 0.674 versus 0.806, two-sided *t*-test $P = 1.2 \times 10^{-16}$), which suggests that these techniques are more effective for picking up genes with a de novo mechanism of disease, compared to recessive genes identified by linkage methods. **c**, Similar to Fig. 5a, the rate ratio is defined by the rate of de novo variants (number per patient) in autism cases divided by the rate in controls. pLoF variants in the most constrained decile of the genome are approximately fourfold more likely to be found in cases compared to controls. **d**, The mean odds ratio of a logistic regression of schizophrenia[28] is plotted for each LOEUF decile. Error bars in **a**–**d** correspond to 95% confidence intervals.



Extended Data Fig. 10 | Applications of constraint metrics to common variant analysis of disease. a, The $\hat{\tau}^*$ coefficient (see Supplementary Information) for each LOEUF decile across 276 independent traits. Unlike the enrichment measure reported in Fig. 5, $\hat{\tau}^*$ is adjusted for 74 baseline genomics annotations. Positive values of $\hat{\tau}^*$ indicate greater per-SNP heritability than would be expected based on the other annotations in the baseline model, whereas negative values indicate depleted per-SNP heritability compared to that baseline expectation. **b**, Enrichment coefficient for each LOEUF decile using different window sizes to define which SNPs to include upstream and downstream of each gene. **c**, Enrichment coefficient for each LOEUF decile across traits after controlling for brain expression and gene size. Results are consistent with those shown in Fig. 5, which indicates that brain gene expression and gene size do not fully explain the enrichment of heritability observed in constrained genes. Error bars represent 95% confidence intervals.

nature research

Corresponding author(s): Konrad Karczewski

Last updated by author(s): Mar 25, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for the collection of data, as this was an opportunistic study.
Data analysis	All code to perform quality control and data analysis is provided in the following Github repos: https://github.com/macarthur-lab/gnomad_qc https://github.com/macarthur-lab/gnomad_lof https://github.com/konradjk/loftee Hail 0.2 is available at: https://hail.is/ Picard version 1.1431 VerifyBamID version 1.0.0 GATK nightly-2015-07-31-g3c929b0, 3.4-89-ge494930, and 3.6-0-g89b7209

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets are described in the manuscript or Supplementary Information, including deposition of the full dataset at https://gnomad.broadinstitute.org, a browser described in the Data Availability section. Data for all figures is available accordingly. There are no restrictions on the aggregate data released.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study was opportunistic, and involved secondary use of all available genome and exome data. No sample size was predetermined. Nevertheless, the current sample size enables the accurate assessment of constraint against pLoF variation for over 72% of genes in the human genome (see Figure 2).
Data exclusions	Sample QC and variant QC for gnomAD are described extensively in the supplementary methods. Notably, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded, as previously established and described [Lek et al., 2016].
Replication	We did not attempt to reproduce any findings in a separate dataset, as no other data set of comparable size exists.
Randomization	As this was a population-based study, and not a case-control study, no randomization was performed.
Blinding	As this was a population-based study, and not a case-control study, blinding was not relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	As an opportunistic collection of data, the participants in this study were not selected based on age, gender, or genotypic information. As described above, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded. The populations are provided in Supplementary Table 7, and there are 64,754 females and 76,702 males. These data were obtained primarily from case-control studies of adult-onset common diseases, including cardiovascular disease, type 2 diabetes, and psychiatric disorders.
Recruitment	As this was an opportunistic secondary use study, we did not recruit any participants.
Ethics oversight	This study was overseen by the Broad Institute's Office of Research Subject Protection and the Partners Human Research Committee, and was given a determination of Not Human Subjects Research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

A structural variation reference for medical and population genetics

https://doi.org/10.1038/s41586-020-2287-8

Received: 2 March 2019

Accepted: 31 March 2020

Published online: 27 May 2020

Open access

Check for updates

Ryan L. Collins[1,2,3,157], Harrison Brand[1,2,4,157], Konrad J. Karczewski[1,5], Xuefang Zhao[1,2,4], Jessica Alföldi[1,5], Laurent C. Francioli[1,5,6], Amit V. Khera[1,2], Chelsea Lowther[1,2,4], Laura D. Gauthier[1,7], Harold Wang[1,2], Nicholas A. Watts[1,5], Matthew Solomonson[1,5], Anne O'Donnell-Luria[1,5], Alexander Baumann[7], Ruchi Munshi[7], Mark Walker[1,7], Christopher W. Whelan[7], Yongqing Huang[7], Ted Brookings[7], Ted Sharpe[7], Matthew R. Stone[1,2], Elise Valkanas[1,2,3], Jack Fu[1,2,4], Grace Tiao[1,5], Kristen M. Laricchia[1,5], Valentin Ruano-Rubio[7], Christine Stevens[1], Namrata Gupta[1], Caroline Cusick[1], Lauren Margolin[1], Genome Aggregation Database Production Team*, Genome Aggregation Database Consortium*, Kent D. Taylor[8], Henry J. Lin[8], Stephen S. Rich[9], Wendy S. Post[10], Yii-Der Ida Chen[8], Jerome I. Rotter[8], Chad Nusbaum[1,154], Anthony Philippakis[7], Eric Lander[1,11,12], Stacey Gabriel[1], Benjamin M. Neale[1,2,5,13], Sekar Kathiresan[1,2,6,14], Mark J. Daly[1,2,5,13], Eric Banks[7], Daniel G. MacArthur[1,2,5,6,155,156] & Michael E. Talkowski[1,2,4,13] ✉

Structural variants (SVs) rearrange large segments of DNA[1] and can have profound consequences in evolution and human disease[2,3]. As national biobanks, disease-association studies, and clinical genetic testing have grown increasingly reliant on genome sequencing, population references such as the Genome Aggregation Database (gnomAD)[4] have become integral in the interpretation of single-nucleotide variants (SNVs)[5]. However, there are no reference maps of SVs from high-coverage genome sequencing comparable to those for SNVs. Here we present a reference of sequence-resolved SVs constructed from 14,891 genomes across diverse global populations (54% non-European) in gnomAD. We discovered a rich and complex landscape of 433,371 SVs, from which we estimate that SVs are responsible for 25–29% of all rare protein-truncating events per genome. We found strong correlations between natural selection against damaging SNVs and rare SVs that disrupt or duplicate protein-coding sequence, which suggests that genes that are highly intolerant to loss-of-function are also sensitive to increased dosage[6]. We also uncovered modest selection against noncoding SVs in *cis*-regulatory elements, although selection against protein-truncating SVs was stronger than all noncoding effects. Finally, we identified very large (over one megabase), rare SVs in 3.9% of samples, and estimate that 0.13% of individuals may carry an SV that meets the existing criteria for clinically important incidental findings[7]. This SV resource is freely distributed via the gnomAD browser[8] and will have broad utility in population genetics, disease-association studies, and diagnostic screening.

SVs are DNA rearrangements that involve at least 50 nucleotides[1]. By virtue of their size and abundance, SVs represent an important mutational force that shape genome evolution and function[2,3], and contribute to germline and somatic diseases[9–11]. The profound effect of SVs is also attributable to the numerous mechanisms by which they can disrupt protein-coding genes and *cis*-regulatory architecture[12]. SVs can be grouped into mutational classes that include 'unbalanced' gains or losses of DNA (for example, copy-number variants, CNVs), and 'balanced' rearrangements that occur without corresponding dosage alterations (such as inversions and translocations)[1] (Fig. 1a). Other common forms of SVs include mobile elements that insert themselves throughout the genome, and multiallelic CNVs (MCNVs) that can exist at high copy numbers[1]. More recently, exotic species of complex SVs have been discovered that involve two or more distinct SV signatures in a single mutational event interleaved on the same allele, and can range from CNV-flanked inversions to rare instances of localized chromosome shattering, such as chromothripsis[13,14]. The diversity of SVs in humans is therefore far greater than has been widely appreciated, as is their influence on genome structure and function.

Although SVs alter more nucleotides per genome than SNVs and short insertion/deletion variants (indels; <50 bp)[1], surprisingly little is known about their mutational spectra on a global scale. The largest published population study of SVs using whole-genome sequencing (WGS) remains the 1000 Genomes Project (n = 2,504; 7× sequence

Lists of affiliations and consortium members appear at the end of the paper.

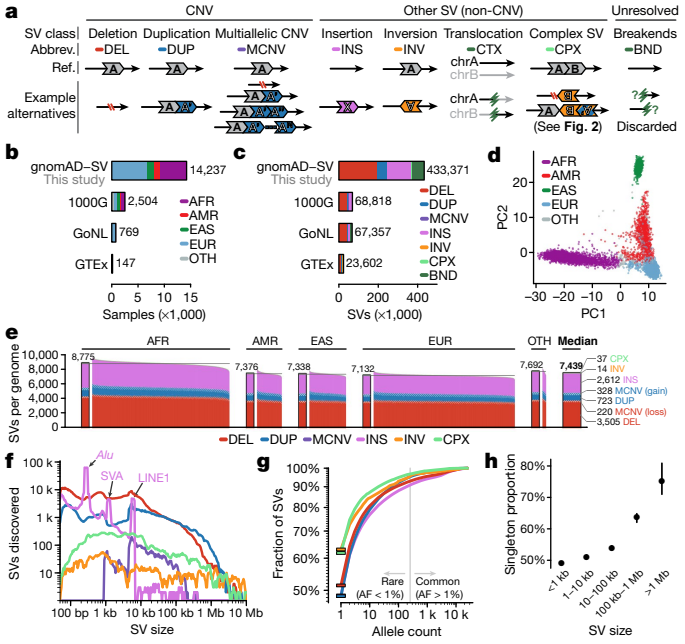


Fig. 1 | Properties of SVs across human populations. a, SV classes catalogued in this study. We also documented unresolved non-reference 'breakends' (BNDs), but they were excluded from all analyses as low-quality variants. **b**, After quality control, we analysed 14,237 samples across continental populations, including African/African American (AFR), Latino (AMR), East Asian (EAS), and European (EUR), or other populations (OTH). Three publicly available WGS-based SV datasets are provided for comparison (1000 Genomes Project (1000G), approximately 7× coverage; Genome of the Netherlands Project (GoNL), around 13× coverage; Genotype-Tissue Expression Project (GTEx), approximately 50× coverage)[1,16,17]. **c**, We discovered 433,371 SVs, and provide counts from previous studies for comparison[1,16,17]. **d**, A principal component (PC) analysis of genotypes for 15,395 common SVs separated samples along axes corresponding to genetic ancestry. **e**, The median genome contained 7,439 SVs. **f**, Most SVs were small. Expected *Alu*, SVA and LINE1 mobile element insertion peaks are marked at approximately 300 bp, 2.1 kb and 6 kb, respectively. **g**, Most SVs were rare (allele frequency (AF) <1%), and 49.8% of SVs were singletons (solid bars). **h**, Allele frequencies were inversely correlated with SV size across all 335,470 resolved SVs in unrelated individuals. Values are mean and 95% confidence interval from 100-fold bootstrapping. Colour codes are consistent between **a**, **c**, **e**–**h**, and between **b** and **d**.

coverage)[1], and the substantial technical challenges of SV discovery from WGS[15] has led to non-uniform SV analyses across contemporary studies[16–20]. Moreover, short-read WGS is unable to capture a subset of SVs accessible to more expensive niche technologies, such as long-read WGS[21]. Owing to the combination of these challenges, SV references are dwarfed by contemporary resources for short variants, such as the Exome Aggregation Consortium (ExAC) and its successor, the Genome Aggregation Database (gnomAD), which have jointly analysed more than 140,000 individuals[4,6]. Publicly available resources such as ExAC and gnomAD have transformed many aspects of human genetics, including defining sets of genes constrained against damaging coding mutations[6] and providing frequency filters for variant interpretation[5]. As short-read WGS is rapidly becoming the predominant technology in large-scale human disease studies, and will probably displace conventional methods for diagnostic screening, there is a mounting need for comparable references of SVs across global populations.

In this study, we developed gnomAD-SV, a sequence-resolved reference for SVs from 14,891 genomes. Our analyses revealed diverse mutational patterns among SVs, and principles of selection acting against reciprocal dosage changes in genes and noncoding *cis*-regulatory elements. From these analyses, we determined that SVs represent more than 25% of all rare protein-truncating events per genome, emphasizing the unrealized potential of routine SV detection in WGS studies. This SV reference has been integrated into the gnomAD browser (http://gnomad.broadinstitute.org) with no restrictions on reuse so that it can be mined for new insights into genome biology and applied as a resource to interpret SVs in diagnostic screening.

SV discovery and genotyping

We analysed WGS data for 14,891 samples (average coverage of 32×) aggregated from large-scale sequencing projects, of which 14,237 (95.6%) passed all quality thresholds, representing a general adult population depleted for severe Mendelian diseases (median age of 49 years) (Supplementary Table 1, Supplementary Figs. 1, 2). This cohort included 46.1% European, 34.9% African or African American, 9.2% East Asian, and 8.7% Latino samples, as well as 1.2% samples from admixed or other populations (Fig. 1). Following family-based analyses using 970 parent–child trios for quality assessments, we pruned all first-degree relatives from the cohort, retaining 12,653 unrelated genomes for subsequent analyses.

We discovered and genotyped SVs using a cloud-based, multi-algorithm pipeline for short-read WGS (Supplementary Fig. 3), which we prototyped in a study of 519 autism quartet families[20]. This pipeline integrated four orthogonal evidence types to capture SVs across the size and allele frequency spectra, including six classes of canonical SVs (Fig. 1a) and 11 subclasses of complex SVs[22] (Fig. 2). We augmented this pipeline with new methods to account for the technical heterogeneity of aggregated datasets (Extended Data Fig. 1, Supplementary Figs. 4, 5), and discovered 433,371 SVs (Fig. 1c). After excluding low-quality SVs, which were predominantly (61.6%) composed of incompletely resolved breakpoint junctions (that is, 'breakends') that lack interpretable alternative allele structures for functional annotation and produce high false-discovery rates[20] (Extended Data Fig. 2a), we retained 335,470 high-quality SVs for subsequent analyses (Supplementary Table 3). This final set of high-quality SVs corresponded to a median of 7,439 SVs per genome, or more than twice the number of variants per genome captured by previous WGS-based SV studies such as the 1000 Genomes Project (3,441 SVs per genome from approximately 7× coverage WGS), which underscores the benefits of high-coverage WGS and improved multi-algorithm ensemble methods for SV discovery.

Given that there are no gold-standard benchmarking procedures for SVs from WGS, we evaluated the technical qualities of gnomAD-SV using seven orthogonal approaches. These analyses are described in detail in Extended Data Figs. 2, 3, Supplementary Figs. 6–12, Supplementary Table 4 and Supplementary Note 1, but we highlight just a few here to demonstrate that gnomAD-SV conforms to many fundamental principles of population genetics, including Mendelian segregation, genotype distributions, and linkage disequilibrium. We found that the precision of gnomAD-SV was comparable to our previous study of 519 autism quartets that attained a 97% molecular validation rate for all de novo SV predictions[20]: in gnomAD, analyses of 970 parent–child trios indicated a median Mendelian violation rate of 3.8% and a heterozygous de novo rate of 3.0%. We also observed that 86% of SVs were in Hardy–Weinberg equilibrium, and common SVs were in strong linkage disequilibrium with nearby SNVs or indels (median peak $R^2 = 0.85$). We performed extensive in silico confirmation of 19,316 SVs predicted from short-read WGS using matched long-read WGS from four samples[21,23], finding a 94.0% confirmation rate with breakpoint-level read evidence, and revealing that 59.8% of breakpoint coordinates were accurate within a single nucleotide of the long-read data. These and other benchmarking approaches suggested that gnomAD-SV was sufficiently sensitive and specific to be used as a reference dataset for most applications in human genomics.

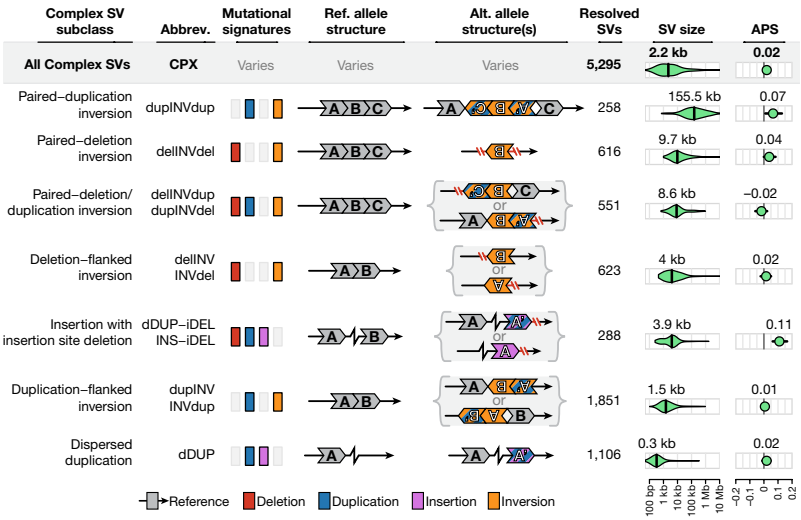


Fig. 2 | Complex SVs are abundant in the human genome. We resolved 5,295 complex SVs across 11 mutational subclasses, 73.7% of which involved at least one inversion. Each subclass is detailed here, including their mutational signatures, structures, abundance, density of SV sizes (vertical line indicates median size), and allele frequencies. Five pairs of subclasses have been collapsed into single rows due to mirrored or similar alternative allele structures (for example, delINV versus INVdel). Two complex SVs did not conform to any subclass (Extended Data Fig. 8).

Population genetics and genome biology

The distribution of SVs across samples matched expectations based on human demographic history, with the top three components of genetic variance separating continental populations (Fig. 1d, Supplementary Fig. 13). African and African American samples exhibited the greatest genetic diversity and their common SVs were in weaker linkage disequilibrium with nearby short variants than Europeans, whereas East Asians featured the highest levels of homozygosity (Fig. 1e, Extended Data Fig. 4a–d, Supplementary Fig. 7). The mutational diversity of gnomAD-SV was extensive: we completely resolved 5,295 complex SVs across 11 mutational subclasses, of which 3,901 (73.7%) involved inverted segments (Fig. 2), confirming that inversion variation is predominantly composed of complex SVs rather than canonical inversions[1,24]. Across all SV classes, most SVs were small (median size of 331 bp) and rare (allele frequency < 1%; 92% of SVs), with half of all SVs (49.8%) appearing as 'singletons' (that is, only one allele observed across all samples) (Fig. 1f, g). Although the proportion of singletons varied by SV class, it was strongly dependent on SV size across all classes, which suggests that the amount of DNA rearranged is a key determinant of selection against most SVs (Fig. 1h, Extended Data Fig. 5a).

Mutation rate estimates for SVs have remained elusive owing to limited sample sizes, poor resolution of conventional technologies, technical challenges of SV discovery, and use of cell line-derived DNA in population studies[1,25]. Here, we used the Watterson estimator[26] to project a mean mutation rate of 0.29 de novo SVs (95% confidence interval 0.13–0.44) per generation in regions of the genome accessible to short-read WGS, or roughly one new SV every 2–8 live births, with mutation rates varying markedly by SV class (Fig. 3a). Although this imperfect method extrapolates from data pooled across unrelated individuals, we previously demonstrated comparable rates from molecularly validated observations in 519 quartet families[20]. Like mutation rates, the distribution of SVs throughout the genome was non-uniform, significantly correlated with repetitive sequence contexts, and was enriched near centromeres and telomeres[23] (Supplementary Fig. 16). These trends were dependent on SV class, as biallelic deletions and duplications were predominantly enriched at telomeres, whereas MCNVs were enriched in centromeric segmental duplications (Fig. 3b–d). Given the reduced sensitivity of short-read WGS in repetitive sequences, this study certainly underestimates the true SV mutation rates; nevertheless, these analyses implicate several aspects of chromosomal context and SV class in determining SV mutation rates throughout the genome.

Dosage sensitivity of coding and noncoding loci

Owing to their size and mutational diversity, SVs can have varied consequences on protein-coding genes[12] (Fig. 4a, Supplementary Fig. 17). In principle, any SV can result in predicted loss-of-function (pLoF), either by deleting coding nucleotides or altering open-reading frames. Coding duplications can result in copy-gain of entire genes, or of a subset of exons within a gene (referred to here as intragenic exonic

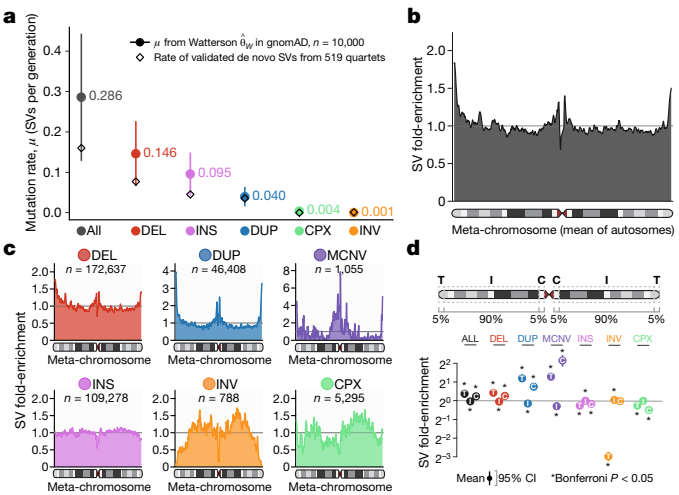


Fig. 3 | Genome-wide mutational patterns of SVs. a, Mutation rates (μ) from the Watterson estimator for each SV class[26]. Bars represent 95% confidence intervals. Rates of molecularly validated de novo SVs from 519 quartet families are provided for comparison[20]. **b**, Smoothed enrichment of SVs per 100-kb window across the average of all autosomes normalized by chromosome arm length (a 'meta-chromosome') (Supplementary Fig. 16). **c**, The distribution of SVs along the meta-chromosome was dependent on variant class. **d**, SV enrichment by class and chromosomal position provided as mean and 95% confidence intervals (CI). C, centromeric; I, interstitial; T, telomeric. *P* values were computed using a two-sided *t*-test and were Bonferroni-adjusted for 21 comparisons. $^*P \le 2.38 \times 10^{-3}$.

duplication, or IED). The average genome in gnomAD-SV contained a mean of 179.8 genes altered by biallelic SVs (144.3 pLoF, 24.3 copy-gain, and 11.2 IED), of which 11.6 were predicted to be completely inactivated by homozygous pLoF (Fig. 4b, Extended Data Fig. 4e–h). When restricted to rare (allele frequency < 1%) SVs, we observed a mean of 10.2 altered genes per genome (5.5 pLoF, 3.4 copy-gain, and 1.3 IED). By comparison, a companion gnomAD paper estimated 122.4 pLoF short variants per genome, of which 16.3 were rare[4]. These analyses suggest that 29.4% of rare heterozygous gene inactivation events per individual are contributed by SVs, or conservatively 25.2% of pLoF events if we exclude IEDs given the context-dependence of their functional impact.

A fundamental question in human genetics is the degree to which natural selection acts on coding and noncoding loci. The proportion of singleton variants has been established as a proxy for strength of selection[6]; however, this metric is confounded for SVs given the strong correlation between allele frequency and SV size, among other factors. Therefore, we developed a new metric, adjusted proportion of singletons (APS), to account for SV class, size, genomic context, and other technical covariates (Extended Data Fig. 5, Supplementary Fig. 14). Under this normalized APS metric, a value of zero corresponds to a singleton proportion comparable to intergenic SVs, whereas values greater than zero reflect purifying selection, similar to the 'mutability-adjusted proportion of singletons' (MAPS) metric used for SNVs[6]. Applying this APS model revealed signals of pervasive selection against nearly all classes of SVs that overlap genes, including intronic SVs, whole-gene inversions, SVs in gene promoters, and deletions as small as a single exon (Fig. 4c, Extended Data Fig. 6, Supplementary Fig. 18). The one notable exception was copy-gain duplications, which showed no clear evidence of selection beyond what could already be explained by their sizes, which were vastly larger than non-copy-gain duplications (median copy-gain duplication size = 134.8 kb; median non-copy-gain duplication size = 2.7 kb; one-tailed Wilcoxon test, $W = 1.18 \times 10^8$, $P < 10^{-100}$). This result could have numerous explanations, but it is consistent with the known diverse evolutionary roles of gene duplication events, including positive selection reported in humans[27,28].

Methods that quantify evolutionary constraint on a per-gene basis, such as the probability of intolerance to heterozygous pLoF variation (pLI)[6] and the pLoF observed/expected upper fraction (LOEUF)[4], have become core resources in human genetics. Nearly all existing metrics, including pLI and LOEUF, are derived from SNVs. Although previous studies have attempted to compute similar scores using large CNVs detected by microarray and exome sequencing[29,30], or to correlate deletions with pLI[18], no gene-level metrics comparable to LOEUF exist for SVs at WGS resolution. To gain insight into this problem, we built a model to estimate the depletion of rare SVs per gene compared to expectations based on gene length, genomic context, and the structure of exons and introns. This model is imperfect, as current sample sizes are too sparse to derive precise gene-level metrics of constraint from SVs. Nevertheless, we found strong concordance between the depletion of rare pLoF SVs and existing pLoF and missense SNV constraint metrics[4] (pLoF Spearman correlation test, $\rho = 0.90$, $P < 10^{-100}$) (Fig. 4d, Supplementary Fig. 19). Notably, a comparable positive correlation was also observed for copy-gain SVs and SNV constraint (pLoF Spearman correlation test, $\rho = 0.78$, $P < 10^{-100}$), whereas a weaker yet significant correlation was detected for IEDs (pLoF Spearman correlation test, $\rho = 0.58$, $P = 2.0 \times 10^{-11}$). As orthogonal support for these trends, we identified an inverse correlation between APS and SNV constraint across all functional categories of SVs, which was consistent with our observed depletion of rare, functional SVs in constrained genes (Extended Data Fig. 6f). These comparisons confirm that selection against most classes of gene-altering SVs mirrors patterns observed for short variants[18,30]. They further suggest that SNV-derived constraint metrics such as LOEUF capture a general correspondence between haploinsufficiency and triplosensitivity for a large fraction of genes in the genome. It therefore appears that the most highly pLoF-constrained genes not only are sensitive to pLoF, but also are more likely to be intolerant to increased dosage and other functional alterations.

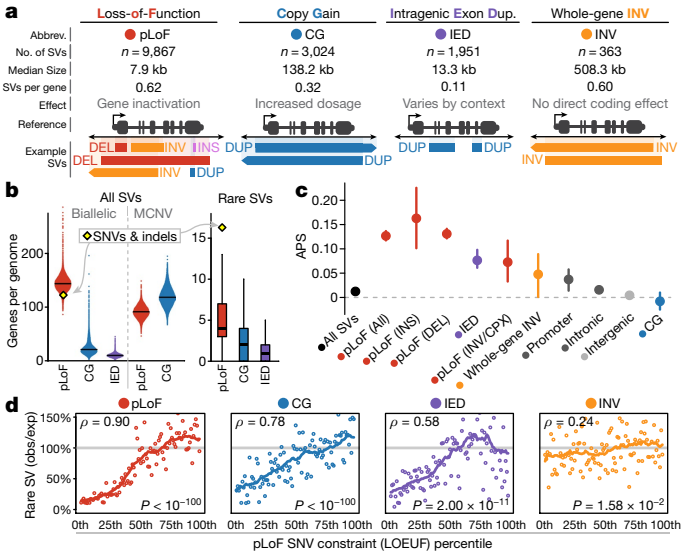


Fig. 4 | Pervasive selection against SVs in genes mirrors coding short variants. a, Four categories of gene-overlapping SVs, with counts of total SVs, median SV size, and mean SVs per gene in gnomAD-SV. **b**, Count of genes altered by SVs per genome. Horizontal lines indicate medians. Sample sizes per category listed in Supplementary Table 9. **c**, APS value for SVs overlapping genes. Bars indicate 100-fold bootstrapped 95% confidence intervals. SVs per category listed in Supplementary Table 9. **d**, Relationships of constraint against pLoF SNVs versus gene-overlapping SVs in 100 bins of around 175 genes each, ranked by SNV constraint[4]. Correlations were assessed with a two-sided Spearman correlation test. Solid lines represent 21-point rolling means. See Supplementary Fig. 19 for comparisons to missense constraint.

In contrast to the well-studied effects of coding variation, the effects of noncoding SVs on regulatory elements are largely unknown. There are a handful of examples of SVs with strong noncoding effects, although they are scarce in humans and model organisms[31,32]. In gnomAD-SV, we explored noncoding dosage sensitivity across 14 regulatory element classes, ranging from high-confidence experimentally validated enhancers to large databases of computationally predicted elements (Supplementary Table 5). We found that noncoding CNVs overlapping most element classes had increased proportions of singletons, although none exceeded the APS observed for pLoF SVs (Fig. 5a). In general, the effects of noncoding deletions appeared stronger than noncoding duplications, and CNVs predicted to delete or duplicate entire elements were under stronger selection than partial element disruption (Fig. 5b). We also observed that primary sequence conservation was correlated with selection against noncoding CNVs (Fig. 5c, d), which provides a foothold for future work on interpretation and functional effect prediction for noncoding SVs. Broadly, these results followed trends we observed for protein-coding SVs, which we interpreted as evidence for weak but widespread selection against CNVs altering most classes of annotated regulatory elements.

Trait association and clinical genetics

Most large-scale trait association studies have only considered SNVs in genome-wide association studies (GWAS). Taking advantage of the sample size and resolution of gnomAD-SV, we evaluated whether SNVs associated with human traits might be in linkage disequilibrium with SVs not directly genotyped in GWAS. We identified 15,634 common SVs (allele frequency >1%) in strong linkage disequilibrium ($R^2 \geq 0.8$) with at least one common short variant (Supplementary

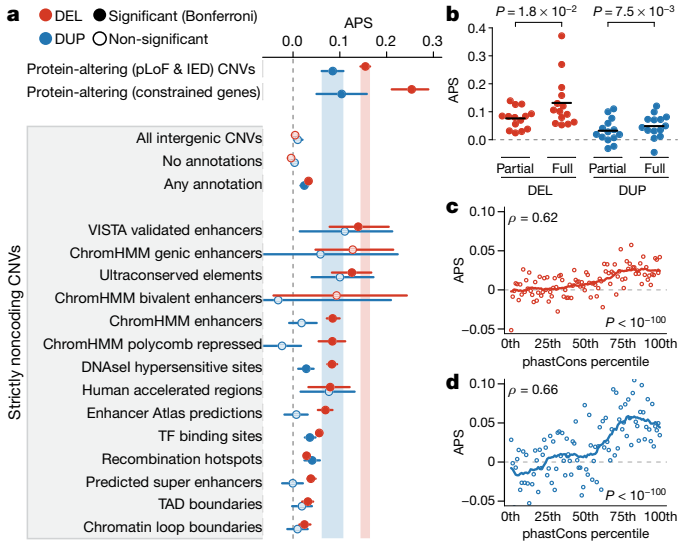


Fig. 5 | Dosage sensitivity in the noncoding genome. a, Strength of selection (APS) for noncoding CNVs overlapping 14 categories of noncoding elements (Supplementary Table 5). Bars reflect 95% confidence intervals from 100-fold bootstrapping. Each category was compared to neutral variation (APS = 0) using a one-tailed *t*-test. Categories surpassing Bonferroni-corrected significance for 32 comparisons are indicated with dark shaded points. SVs per category listed in Supplementary Table 9. DEL, deletion; DUP, duplication; TAD, topologically associating domain; TF, transcription factor. **b**, CNVs that completely covered elements ('full') had significantly higher average APS values than CNVs that only partially covered elements ('partial'). *P* values calculated using a two-tailed paired two-sample *t*-test for the 14 categories from **a**. **c**, **d**, Spearman correlations between sequence conservation and APS for noncoding deletions ($n = 143{,}353$) (**c**) and duplications ($n = 30{,}052$) (**d**). Noncoding CNVs were sorted into 100-percentile bins based on the sum of the phastCons scores overlapped by the CNV. Correlations were assessed with a two-sided Spearman correlation test. Solid lines represent 21-point rolling means.

Fig. 7), 14.8% of which matched a reported association from the NHGRI-EBI GWAS catalogue or a recent analysis of 4,203 phenotypes in the UK Biobank[33,34]. Common SVs in linkage disequilibrium with GWAS variants were enriched for genic SVs across multiple functional categories (Supplementary Table 6), and included candidate SVs such as a deletion of a thyroid enhancer in the first intron of *ATP6V0D1* at a hypothyroidism-associated locus[34] (Extended Data Fig. 7). We also identified matches for previously proposed causal SVs tagged by common SNVs, including pLoF deletions of *CFHR3* or *CFHR1* in nephropathies and of *LCE3B* or *LCE3C* in psoriasis[35,36]. These results demonstrate the value of imputing SVs into GWAS, and for the eventual unification of short variants and SVs in all trait association studies. Given the potential value of this resource, we have released these linkage disequilibrium maps in Supplementary Table 7.

As genomic medicine advances towards diagnostic screening at sequence resolution, computational methods for variant discovery from WGS and population references for interpretation will become indispensable. One category of disease-associated SVs, recurrent CNVs mediated by homologous segmental duplications known as genomic disorders, are particularly important because they collectively represent a common cause of developmental disorders[37]. Accurate detection of large, repeat-mediated CNVs is thus crucial for WGS-based diagnostic testing as chromosomal microarray is the recommended first-tier diagnostic screen at present for unexplained developmental disorders[37]. Using gnomAD-SV, we evaluated our ability to detect genomic disorders in WGS data by calculating CNV carrier frequencies for 49 genomic disorders across 10,047 unrelated samples with no known neuropsychiatric disease and found that CNV carrier frequencies in gnomAD-SV were consistent with those reported from chromosomal microarray in the UK Biobank[38] ($R^2 = 0.669$; Pearson correlation test, $P = 7.38 \times 10^{-13}$) (Fig. 6a, Supplementary Table 8, Supplementary Fig. 20). The frequencies of carriers of genomic disorders did not vary significantly among populations, with the exception of duplications of *NPHP1* at 2q13, in which carrier frequencies in East Asian samples were up to 4.6-fold higher than in other populations, further highlighting the potential for variant interpretation to be confounded by the limited diversity of existing SV references (Supplementary Fig. 21).

In the context of variant interpretation, the current gnomAD-SV resource will permit a screening threshold of allele frequencies less than 0.1% when matching on ancestry to the populations sampled here, and allele frequencies less than 0.004% globally. In the current release, we catalogued at least one pLoF or copy-gain variant for 36.9% and 23.7% of all autosomal genes, respectively, and 490 genes with at least one homozygous pLoF SV (Fig. 6b, Extended Data Fig. 6e, Supplementary Fig. 22). We also benchmarked carrier rates for several categories of clinically relevant variants in gnomAD-SV. First, 0.32% of samples carried a very rare (allele frequency < 0.1%) SV resulting in pLoF of a gene for which incidental findings are clinically actionable, nearly half of which (that is, 0.13% of all samples) would meet diagnostic criteria as pathogenic or likely pathogenic based upon the American College of Medical Genetics (ACMG) recommendations[7] (Fig. 6c). Second, 7.22% of individuals were heterozygous carriers of rare pLoF SVs in known recessive developmental disorder genes[39]. Third, we estimated that 3.8% of the general population (95% confidence interval of 3.2–4.6%) carries at least one very large (≥1 Mb) rare autosomal SV, roughly half of which (45.2%) were balanced or complex (Fig. 6d). Among these was an example of localized chromosome shattering involving at least 49 breakpoints, yet resulting in largely balanced products, reminiscent of chromothripsis, in an adult with no known severe disease or DNA repair defect[13,14,22] (Fig. 6e, Extended Data Fig. 8). Collectively, these analyses highlight the potential of gnomAD-SV and WGS-based SV methods to augment disease-association studies and clinical interpretation across a broad spectrum of variant classes and study designs.

Discussion

Human genetic research and clinical diagnostics are becoming increasingly invested in capturing the complete landscape of variation in individual genomes. Ambitious international initiatives to generate short-read WGS in many thousands of individuals from common disease cohorts have underwritten this goal[40,41], and millions of genomes will be sequenced in the coming years from national biobanks[42,43]. A central challenge to these efforts will be the uniform analysis and interpretation of all variation accessible to WGS, particularly SVs, which are frequently invoked as a source of added value offered by WGS. Indeed, early WGS studies in cardiovascular disease and autism have been largely consistent in their analyses of short variants, but every study has differed in its analysis of SVs[18–20,40,41]. Thus, while ExAC and gnomAD have prompted remarkable advances in medical and population genetics for short variants, the same gains have not yet been realized for SVs. Although gnomAD-SV is not exhaustively comprehensive, it was derived from WGS methods and a reference genome that match those currently used in many research and clinical settings, which will help to facilitate the eventual standardization of SV discovery, analysis, and interpretation across studies.

Most foundational assumptions about human genetic variation were consistent between SVs and short variants in gnomAD, most notably that SVs segregate stably on haplotypes in the population and experience selection commensurate with their predicted biological consequences. This study also spotlights unique aspects of SVs, such as their remarkable mutational diversity, their varied functional effects on coding sequence, and the intense selection against large and complex

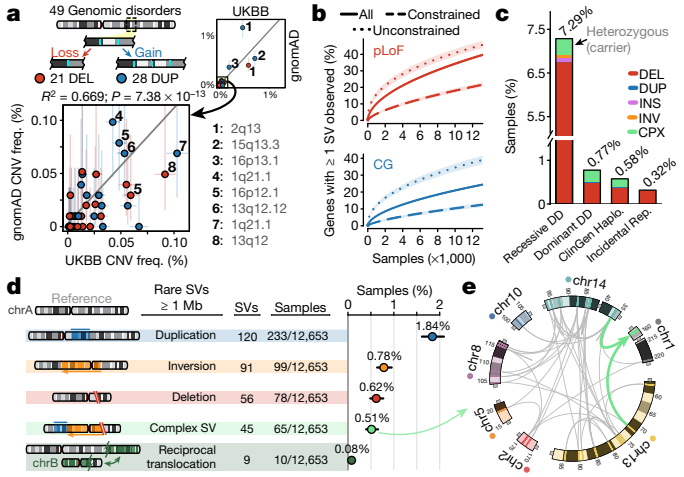


Fig. 6 | gnomAD-SV as a resource for clinical WGS interpretation. **a**, Comparison of carrier frequencies for 49 putatively disease-associated deletions (red) and duplications (blue) at genomic disorder loci between gnomAD-SV and microarray analyses in the UK Biobank (UKBB)[38]. Light bars indicate binomial 95% confidence intervals. Solid grey line represents linear best fit. **b**, At least one pLoF or copy-gain SV was detected in 36.9% and 23.7% of all autosomal genes, respectively. 'Constrained' and 'unconstrained' includes the least and most constrained 15% of all genes based on LOEUF[4], respectively. **c**, Carrier rates for very rare (allele frequency < 0.1%) pLoF SVs in medically relevant genes across several gene lists[7,39,44]. SVs per category listed in Supplementary Table 9. **d**, Carrier rates for very large (≥1 Mb) rare autosomal SVs among 12,653 genomes. Bars represent binomial 95% confidence intervals. **e**, A complex SV involving at least 49 breakpoints and seven chromosomes (also see Extended Data Fig. 8). Teal arrows indicate insertion point into chromosome 1.

SVs. Our analyses also demonstrate that gene-altering effects of SVs beyond pLoF are remarkably similar to the mutational constraints of SNVs, and that SNV constraint metrics are not specific to haploinsufficiency but underlie a general intolerance to alterations of both gene dosage and structure. Beyond genes, we uncovered widespread but modest selection against noncoding dosage alterations of many families of *cis*-regulatory elements. This study represents one of the largest empirical assessments of noncoding dosage sensitivity in humans, and underscores that: (1) few—if any—classes of noncoding *cis*-regulatory variants are likely to experience selection as strong as protein-truncating variants; (2) sequence conservation is unsurprisingly one of the strongest features associated with selection against noncoding SVs; and (3) current WGS sample sizes are vastly underpowered to identify individual constrained functional elements in the noncoding genome.

The value of the multi-algorithm ensemble approach and deep WGS is evident in the improved sensitivity of SV detection in gnomAD-SV. However, short-read WGS remains limited by comparison to emerging long-read technologies[21]. Given that short-read WGS is blind to a disproportionate fraction of repeat-mediated SVs and small insertions by comparison to long-read methods, this study certainly underestimates the true mutation rates within such hypermutable regions. Similarly, although our approach involves extensive methods to resolve complex SV alleles, some variants such as high-copy-state MCNVs often involve complicated haplotype configurations, and we expect that emerging de novo assembly and graph-based genome representations will greatly expand our knowledge of such SVs[21,23]. Nonetheless, 92.7% of all known autosomal protein-coding nucleotides are not localized to simple- or low-copy repeats, and therefore we expect that the catalogues of SVs accessible to short-read WGS across large populations like gnomAD-SV will capture a majority of the most interpretable gene-disrupting SVs in humans.

The scale of short-read WGS datasets currently in production has magnified the need for publicly available SV resources, and gnomAD-SV represents an initial effort to fill this void. Although these data remain insufficient to derive accurate estimates of gene-level constraint, sequence-specific mutation rates, and intolerance to noncoding SVs, they provide a step towards these goals and reinforce the value of data sharing and harmonized analyses of aggregated genomic data sets. These data have been made available without restrictions on reuse (https://gnomad.broadinstitute.org), and this resource will catalyse new discoveries in basic research while providing immediate clinical utility for the interpretation of rare structural rearrangements across human populations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2287-8.

1. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
2. Perry, G. H. et al. Copy number variation and evolution in humans and chimpanzees. *Genome Res.* **18**, 1698–1710 (2008).
3. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
4. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* https://doi.org/10.1038/s41586-020-2308-7 (2020).
5. Walsh, R. et al. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.* **19**, 192–203 (2017).
6. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
7. Green, R. C. et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013).
8. Karczewski, K. J. et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45** (D1), D840–D845 (2017).
9. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
10. Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
11. Talkowski, M. E. et al. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149**, 525–537 (2012).
12. Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet.* **24**, 238–245 (2008).
13. Chiang, C. et al. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat. Genet.* **44**, 390–397 (2012).
14. Kloosterman, W. P. et al. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.* **20**, 1916–1924 (2011).
15. Kosugi, S. et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 117 (2019).
16. Hehir-Kwa, J. Y. et al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7**, 12989 (2016).
17. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
18. Brandler, W. M. et al. Paternally inherited *cis*-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
19. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722.e12 (2017).
20. Werling, D. M. et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).
21. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
22. Collins, R. L. et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* **18**, 36 (2017).
23. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019).
24. Brand, H. et al. Paired-duplication signatures mark cryptic inversions and other complex structural variation. *Am. J. Hum. Genet.* **97**, 170–176 (2015).
25. Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
26. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
27. Dennis, M. Y. & Eichler, E. E. Human adaptation and evolution by segmental duplication. *Curr. Opin. Genet. Dev.* **41**, 44–52 (2016).

28. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3**, research0008.1 (2002).
29. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154 (2010).
30. Ruderfer, D. M. et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat. Genet.* **48**, 1107–1111 (2016).
31. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
32. Redin, C. et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* **49**, 36–45 (2016).
33. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47** (D1), D1005–D1012 (2019).
34. Neale, B. UK Biobank GWAS round 2; http://www.nealelab.is/uk-biobank/ (2018).
35. Xie, J. et al. Fine mapping implicates a deletion of *CFHR1* and *CFHR3* in protection from IgA nephropathy in Han Chinese. *J. Am. Soc. Nephrol.* **27**, 3187–3194 (2016).
36. de Cid, R. et al. Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41**, 211–215 (2009).
37. Miller, D. T. et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **86**, 749–764 (2010).
38. Owen, D. et al. Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. *BMC Genomics* **19**, 867 (2018).
39. Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
40. Choi, S. H. et al. Association between titin loss-of-function variants and early-onset atrial fibrillation. *J. Am. Med. Assoc.* **320**, 2354–2364 (2018).
41. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391 (2018).
42. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).
43. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
44. Rehm, H. L. et al. ClinGen—the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
45. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
46. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
47. Zhao, X., Weber, A. M. & Mills, R. E. A recurrence-based approach for validating structural variation using long-read sequencing technology. *Gigascience* **6**, 1–9 (2017).
48. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
49. Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [3]Division of Medical Sciences, Harvard Medical School, Boston, MA, USA. [4]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [5]Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [6]Department of Medicine, Harvard Medical School, Boston, MA, USA. [7]Data Science Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA. [9]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [10]Johns Hopkins University School of Medicine, Baltimore, MD, USA. [11]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [12]Department of Biology, MIT, Cambridge, MA, USA. [13]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [14]Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA. [154]Present address: Cellarity Inc., Cambridge, MA, USA. [155]Present address: Centre for Population Genomics, Garvan Institute of Medical Research, and UNSW Sydney, Sydney, Australia. [156]Present address: Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia. [157]These authors contributed equally: Ryan L. Collins, Harrison Brand. ✉e-mail: talkowsk@broadinstitute.org

Genome Aggregation Database Production Team

Jessica Alföldi[1,15], Irina M. Armean[1,15,16], Eric Banks[7], Louis Bergelson[7], Kristian Cibulskis[7], Ryan L. Collins[1,2,3], Kristen M. Connolly[17], Miguel Covarrubias[7], Beryl Cummings[1,3,15], Mark J. Daly[1,13,15], Stacey Donnelly[1], Yossi Farjoun[7], Steven Ferriera[18], Laurent Francioli[1,15], Stacey Gabriel[18], Laura D. Gauthier[7], Jeff Gentry[7], Namrata Gupta[1,18], Thibault Jeandet[7], Diane Kaplan[7], Konrad J. Karczewski[1,15], Kristen M. Laricchia[1,15], Christopher Llanwarne[7], Eric V. Minikel[1], Ruchi Munshi[7], Benjamin M. Neale[1,15], Sam Novod[7], Anne H. O'Donnell-Luria[1,19,20], Nikelle Petrillo[7], Timothy Poterba[1,13,15], David Roazen[7], Valentin Ruano-Rubio[7], Andrea Saltzman[1], Kaitlin E. Samocha[21], Molly Schleicher[1], Cotton Seed[13,15], Matthew Solomonson[1,15], Jose Soto[7], Grace Tiao[1,15], Kathleen Tibbetts[7], Charlotte Tolonen[7], Christopher Vittal[13,15], Gordon Wade[7], Arcturus Wang[1,13,15], Qingbo Wang[1,3,15], James S. Ware[1,22,23], Nicholas A. Watts[1,15], Ben Weisburd[7] & Nicola Whiffin[1,22,23]

[15]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [16]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [17]Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [18]Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [19]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [20]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [21]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [22]National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. [23]Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London, UK.

Genome Aggregation Database Consortium

Carlos A. Aguilar Salinas[24], Tariq Ahmad[25], Christine M. Albert[26,27], Diego Ardissino[28], Gil Atzmon[29,30,31], John Barnard[32], Laurent Beaugerie[33], Emelia J. Benjamin[34,35,36], Michael Boehnke[37], Lori L. Bonnycastle[38], Erwin P. Bottinger[39], Donald W. Bowden[40,41,42], Matthew J. Bown[43,44], John C. Chambers[45,46,47], Juliana C. Chan[48], Daniel Chasman[6,26], Judy Cho[39], Mina K. Chung[32], Bruce Cohen[6,49], Adolfo Correa[50], Dana Dabelea[51], Mark J. Daly[1,13,15], Dawood Darbar[52,53], Ravindranath Duggirala[54], Josée Dupuis[34,55], Patrick T. Ellinor[1,56], Roberto Elosua[57,58,59], Jeanette Erdmann[60,61,62], Tõnu Esko[1,63], Martti Färkkilä[64], Jose Florez[1,7,65,66,67], Andre Franke[68], Gad Getz[6,69,70,71,72], Benjamin Glaser[73], Stephen J. Glatt[74], David Goldstein[75,76], Clicerio Gonzalez[77], Leif Groop[78,79], Christopher Haiman[80], Craig Hanis[81], Matthew Harms[82,83], Mikko Hiltunen[84], Matti M. Holi[85], Christina M. Hultman[86,87], Mikko Kallela[88], Jaakko Kaprio[79,89], Sekar Kathiresan[1,2,6,90], Bong-Jo Kim[91], Young Jin Kim[91], George Kirov[92], Jaspal Kooner[22,46,47], Seppo Koskinen[93], Harlan M. Krumholz[94], Subra Kugathasan[95], Soo Heon Kwak[96], Markku Laakso[97,98], Terho Lehtimäki[99], Ruth J. F. Loos[39,100], Steven A. Lubitz[1,56], Ronald C. W. Ma[48,101,102], Daniel G. MacArthur[1,15], Jaume Marrugat[58,103], Kari M. Mattila[99], Steven McCarroll[13,104], Mark I. McCarthy[105,106,107], Dermot McGovern[108], Ruth McPherson[109], James B. Meigs[1,6,110], Olle Melander[111], Andres Metspalu[63], Benjamin M. Neale[1,15], Peter M. Nilsson[112], Michael C. O'Donovan[92], Dost Ongur[6,49], Lorena Orozco[113], Michael J. Owen[92], Colin N. A. Palmer[114], Aarno Palotie[13,15,79], Kyong Soo Park[96,115], Carlos Pato[116], Ann E. Pulver[117], Nazneen Rahman[118], Anne M. Remes[119], John D. Rioux[120,121], Samuli Ripatti[1,79,89], Dan M. Roden[122,123], Danish Saleheen[124,125,126], Veikko Salomaa[127], Nilesh J. Samani[43,44], Jeremiah Scharf[1,2,13], Heribert Schunkert[128,129], Moore B. Shoemaker[130], Pamela Sklar[131,132,133,158], Hilkka Soininen[134], Harry Sokol[33], Tim Spector[135], Patrick F. Sullivan[86,136], Jaana Suvisaari[127], E. Shyong Tai[137,138,139], Yik Ying Teo[137,140,141], Tuomi Tiinamaija[79,142,143], Ming Tsuang[144,145], Dan Turner[146], Teresa Tusie-Luna[147,148], Erkki Vartiainen[89], James S. Ware[1,22,23], Hugh Watkins[149], Rinse K. Weersma[150], Maija Wessman[79,142], James G. Wilson[151] & Ramnik J. Xavier[152,153]

[24]Unidad de Investigacion de Enfermedades Metabolicas, Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico. [25]Peninsula College of Medicine and Dentistry, Exeter, UK. [26]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA. [27]Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [28]Department of Cardiology, University Hospital, Parma, Italy. [29]Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel. [30]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, USA. [31]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. [32]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [33]Gastroenterology Department, Sorbonne Université, APHP, Saint Antoine Hospital, Paris, France. [34]Framingham Heart Study, National Heart, Lung, & Blood Institute and Boston University, Framingham, MA, USA. [35]Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, USA. [36]Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, USA. [37]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [38]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [39]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [40]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [41]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [42]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [43]Department of Cardiovascular Sciences, University of Leicester, Leicester, UK. [44]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK.

[45]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [46]Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK. [47]Imperial College Healthcare NHS Trust, Imperial College London, London, UK. [48]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. [49]Program for Neuropsychiatric Research, McLean Hospital, Belmont, MA, USA. [50]Department of Medicine, University of Mississippi Medical Center, Jackson, MI, USA. [51]Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA. [52]Department of Pharmacology, University of Illinois at Chicago, Chicago, IL, USA. [53]Department of Medicine, University of Illinois at Chicago, Chicago, IL, USA. [54]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA. [55]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [56]Cardiac Arrhythmia Service, Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [57]Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain. [58]Centro de Investigación en Red en Enfermedades Cardiovasculares (CIBERCV), Barcelona, Catalonia, Spain. [59]Department of Medicine, Medical School, University of Vic-Central University of Catalonia, Vic, Catalonia, Spain. [60]Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany. [61]DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, Lübeck, Germany. [62]University Heart Center, Lübeck, Lübeck, Germany. [63]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [64]Clinic of Gastroenterology, Helsinki University Hospital, Helsinki University, Helsinki, Finland. [65]Diabetes Unit, Massachusetts General Hospital, Boston, MA, USA. [66]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [67]Program in Metabolism, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [68]Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany. [69]Bioinformatics Consortium, Massachusetts General Hospital, Boston, MA, USA. [70]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [71]Cancer Center, Massachusetts General Hospital, Boston, MA, USA. [72]Cancer Genome Computational Analysis Group, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [73]Endocrinology and Metabolism Department, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [74]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA. [75]Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [76]Department of Genetics and Development, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [77]Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Cuernavaca, Mexico. [78]Genomics, Diabetes and Endocrinology, Lund University, Lund, Sweden. [79]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland. [80]Lund University Diabetes Centre, Malmö, Sweden. [81]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA. [82]Department of Neurology, Columbia University, New York, NY, USA. [83]Institute of Genomic Medicine, Columbia University, New York, NY, USA. [84]Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland. [85]Department of Psychiatry, Helsinki University Central Hospital, Lapinlahdentie, Helsinki, Finland. [86]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [87]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [88]Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland. [89]Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [90]Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [91]Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, South Korea. [92]MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Cardiff, USA. [93]Department of Health, National Institute for Health and Welfare (THL), Helsinki, Finland. [94]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, Connecticut Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, CT, USA. [95]Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, GA, USA. [96]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. [97]Institute of Clinical Medicine, The University of Eastern Finland, Kuopio, Finland. [98]Institute of Clinical Medicine Neurology, Kuopio University Hospital, Kuopio, Finland. [99]Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [100]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [101]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [102]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. [103]Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain. [104]Department of Genetics, Harvard Medical School, Boston, MA, USA. [105]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Headington, Oxford, UK. [106]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [107]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [108]F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [109]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada. [110]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [111]Department of Clinical Sciences, University Hospital Malmo Clinical Research Center, Lund University, Malmo, Sweden. [112]Department of Clinical Sciences, Lund University, Skane University Hospital, Malmo, Sweden. [113]Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, Mexico. [114]Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. [115]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea. [116]Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, CA, USA. [117]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MA, USA. [118]Division of Genetics and Epidemiology, Institute of Cancer Research, London, USA. [119]Research Unit of Clinical Neuroscience, University of Oulu, Oulu, Finland. [120]Research Center, Montreal Heart Institute, Montreal, Quebec, Canada. [121]Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada. [122]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [123]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [124]Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [125]Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [126]Center for Non-Communicable Diseases, Karachi, Pakistan. [127]National Institute for Health and Welfare, Helsinki, Finland. [128]Deutsches Herzzentrum München, Munich, Germany. [129]Technische Universität München, Munich, Germany. [130]Division of Cardiovascular Medicine, Nashville VA Medical Center and Vanderbilt University, School of Medicine, Nashville, TN, USA. [131]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [132]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [133]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [134]Institute of Clinical Medicine Neurology, University of Eastern Finland, Kuopio, Finland. [135]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [136]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [137]Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore, Singapore. [138]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [139]Duke-NUS Graduate Medical School, Singapore, Singapore. [140]Life Sciences Institute, National University of Singapore, Singapore, Singapore. [141]Department of Statistics and Applied Probability, National University of, Singapore, Singapore. [142]Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland. [143]HUCH Abdominal Center, Helsinki University Hospital, Helsinki, Finland. [144]Department of Psychiatry, Center for Behavioral Genomics, University of California, San Diego, CA, USA. [145]Institute of Genomic Medicine, University of California, San Diego, CA, USA. [146]Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel. [147]Instituto de Investigaciones Biomédicas UNAM, Mexico City, Mexico. [148]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. [149]Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [150]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. [151]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [152]Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [153]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. [158]Deceased: Pamela Sklar.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All gnomAD-SV site-frequency data for appropriately consented samples (n = 10,847) have been distributed in VCF and BED format via the gnomAD browser (https://gnomad.broadinstitute.org/downloads/), as well as from NCBI dbVar under accession nstd166. Furthermore, these SVs have been integrated directly into the gnomAD browser[8]. The architecture of the gnomAD browser is described in the main gnomAD study[4], as well as instructions for how to access and query the data hosted therein.

Code availability

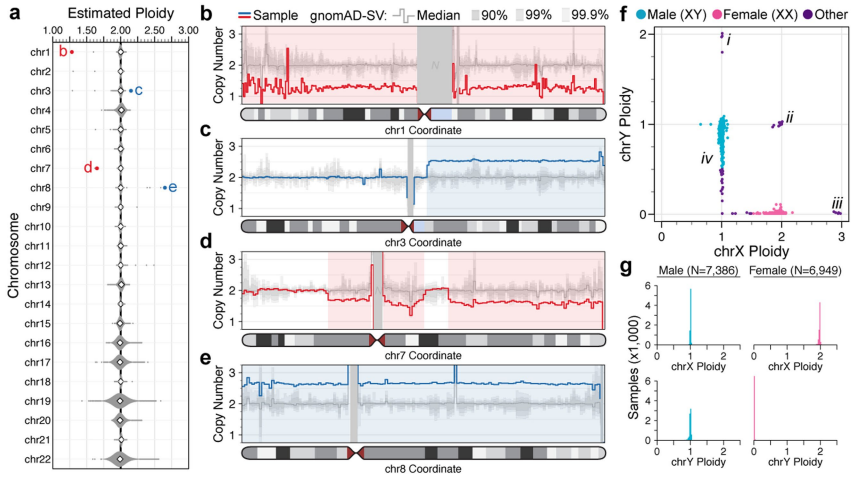
The gnomAD-SV discovery pipeline is publicly available via a series of methods configured for the FireCloud/Terra platform (https://portal.firecloud.org/#methods) under the methods namespace 'Talkowski-SV'. The svtk software package used extensively in the gnomAD-SV discovery pipeline is publicly available via GitHub (https://github.com/talkowski-lab/svtk). Most custom scripts used in the production and/or analysis of the gnomAD-SV dataset are publicly available via GitHub (https://github.com/talkowski-lab/gnomad-sv-pipeline). All code is made available under the MIT license, unless stated otherwise.

Acknowledgements We thank the many individuals whose sequence data are aggregated in gnomAD for their contributions to research, and the users of gnomAD for their collaborative feedback. We are grateful to the families at the participating Simons Simplex Collection (SSC) sites, as well as the SSC principal investigators. We thank T. Hefferon of the NIH National Center for Biotechnology Information for his help hosting gnomAD-SV on dbVar. We have complied with all relevant ethical regulations. Research and contributing authors were supported by resources from the Broad Institute, the National Institutes of Health (NIH) (R01MH115957 to M.E.T., B.N. and D.G.M.; UM1HG008895 to M.J.D., B.N., S.G., E.S.L., S.K., M.E.T.; R01HD081256, P01GM061354, R01HD091797, R01HD096326, R01MH111776, R01HD099547 to M.E.T.; U01MH105669 to M.J.D., B.N. and M.E.T.; P50HD028138 to B.N. and M.E.T.; P01HD068250 to H.B.) and the Simons Foundation for Autism Research Initiative (SFARI #573206 to M.E.T.). R.L.C. was supported by NHGRI T32HG002295 and NSF GRFP #2017240332. H.B. was supported by NIDCR K99DE026824. A.V.K. was supported by NHGRI K08HG010155. M.E.T. was supported by Desmond and Ann Heathwood. MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-00107, and UL1-TR-001420. MESA family is conducted and supported by the NHLBI in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258 and R01HL071259, by the National Center for Research Resources, grant UL1RR033176, and the National Center for Advancing Translational Sciences ULTR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

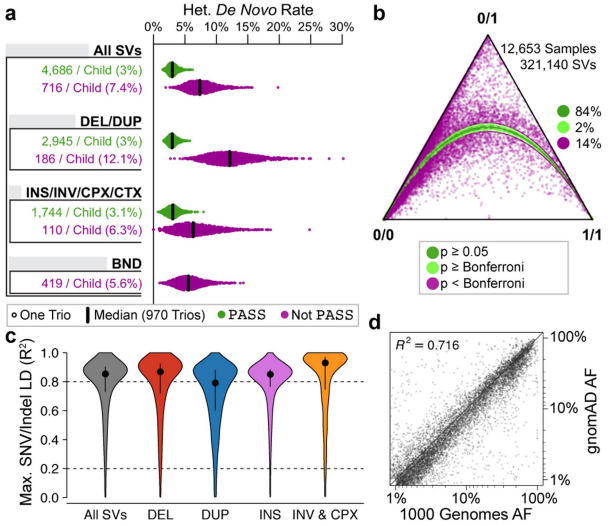
Author contributions R.L.C., H.B., K.J.K., X.Z., J.A., L.C.F, C.L., A.O'D.-L., E.V., H.J.L., J.I.R, M.J.D., D.G.M. and M.E.T. contributed to the writing of the manuscript and generation of figures. R.L.C., H.B., K.J.K., X.Z., L.C.F., C.L., L.D.G., H.W., E.V., J.F., M.J.D., E.B., D.G.M. and M.E.T. contributed to the analysis of data. R.L.C., H.B., X.Z., L.D.G., H.W., N.A.W., M.S., A.B., R.M., M.W., C.W., Y.H., T.B., T.S., M.R.S., E.V., J.F., V.R.-R., C.N., A.P., B.M.N., E.B., D.G.M. and M.E.T. developed tools and methods that enabled the scientific discoveries herein. R.L.C., H.B., K.J.K., X.Z., J.A., L.C.F., A.V.K., L.D.G., H.W., N.A.W., M.S., A.O'D.-L., A.B., R.M., G.T., K.M.L., C.S., N.G., C.C., L.M., K.D.T., H.J.L., S.S.R., W.P., Y.-D.I.C., J.I.R., C.N., A.P., E.L., S.G., B.M.N., S.K., M.J.D., E.B., D.G.M. and M.E.T. contributed to the production and quality control of the gnomAD dataset. All authors listed under The Genome Aggregation Database Consortium contributed to the generation of the primary data incorporated into the gnomAD resource. All authors reviewed the manuscript. R.L.C. and H.B. contributed equally to this study.

Competing interests K.J.K. owns stock in Personalis. A.O'D.-L. has received honoraria from ARUP and Chan Zuckerberg Initiative. B.M.N. is a member of the scientific advisory board at Deep Genomics and consultant for Camp4 Therapeutics, Takeda Pharmaceutical, and Biogen. M.J.D. is a founder of Maze Therapeutics. D.G.M. is a founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme. M.E.T has received research support from Levo Therapeutics. All other authors declare no competing interests. S.K. is an employee of Verve Therapeutics, and holds equity in Verve Therapeutics, Maze Therapeutics, Catabasis, and San Therapeutics. He is a member of the scientific advisory boards for Regeneron Genetics Center and Corvidia Therapeutics; he has served as a consultant for Acceleron, Eli Lilly, Novartis, Merck, Novo Nordisk, Novo Ventures, Ionis, Alnylam, Aegerion, Haug Partners, Noble Insights, Leerink Partners, Bayer Healthcare, Illumina, Color Genomics, MedGenome, Quest, and Medscape; he reports patents related to a method of identifying and treating a person having a predisposition to or afflicted with cardiometabolic disease (20180010185) and a genetics risk predictor (20190017119).

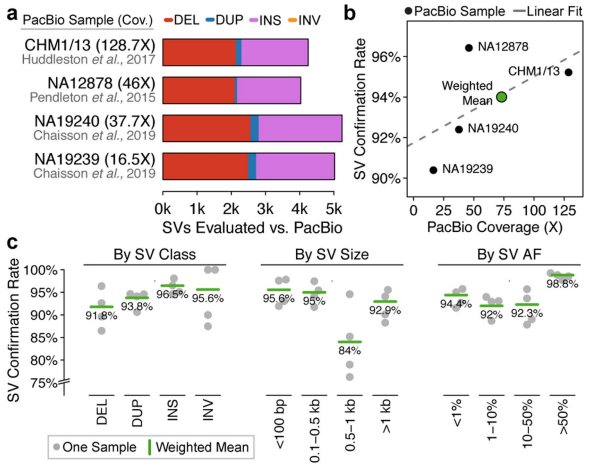
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2287-8.
Correspondence and requests for materials should be addressed to M.E.T.
Peer review information *Nature* thanks Don Conrad, Jan Korbel, Tobias Rausch and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.



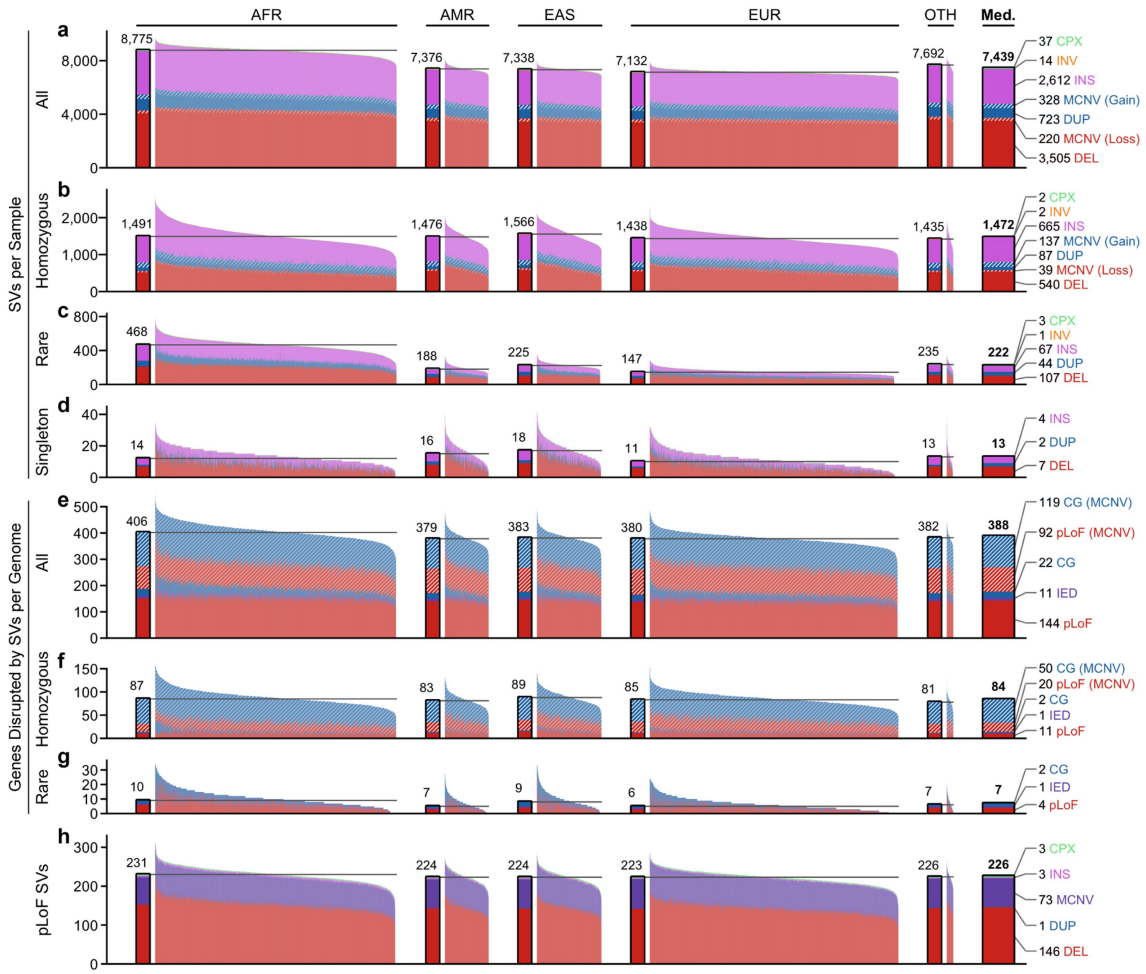
Extended Data Fig. 1 | Detection of chromosome-scale dosage alterations. We estimated ploidy (that is, whole-chromosome copy number) for all 24 chromosomes per sample. **a**, Distribution of autosome ploidy estimates across 14,378 samples passing initial data quality thresholds. White diamonds indicate medians. Individual points are outlier samples at least three standard deviations away from the cohort-wide mean. The outlier points marked in red and blue correspond to the samples highlighted in **b**–**e**. **b**–**e**, Samples with outlier autosome ploidy estimates typically contained somatic or mosaic chromosomal abnormalities, such as somatic aneuploidy of chromosome 1 (chr1) (**b**) or chromosome 8 (**e**), or large focal somatic or mosaic CNVs on chromosome 3 (**c**) and chromosome 7 (**d**). Each panel depicts copy-number estimates in 1-Mb bins for each rearranged sample in red or blue. Dark, medium and light-grey background shading indicates the range of copy number estimates for 90%, 99% and 99.9% of all gnomAD-SV samples, respectively, and the medium grey line indicates the median copy number estimate across all samples. Regions of unalignable N-masked bases >1 Mb in the reference genome are masked with grey rectangles. **f**, Sex chromosome ploidy estimates for all samples from **a**. We inferred karyotypic sex by clustering samples to their nearest integer ploidy for sex chromosomes. Several abnormal sex chromosome ploidies are marked, including XYY (i), XXY (ii), XXX (iii), and mosaic loss-of-Y (iv). **g**, Histogram representation of the data from **f**. Essentially all samples conformed to canonical sex chromosome ploidies.



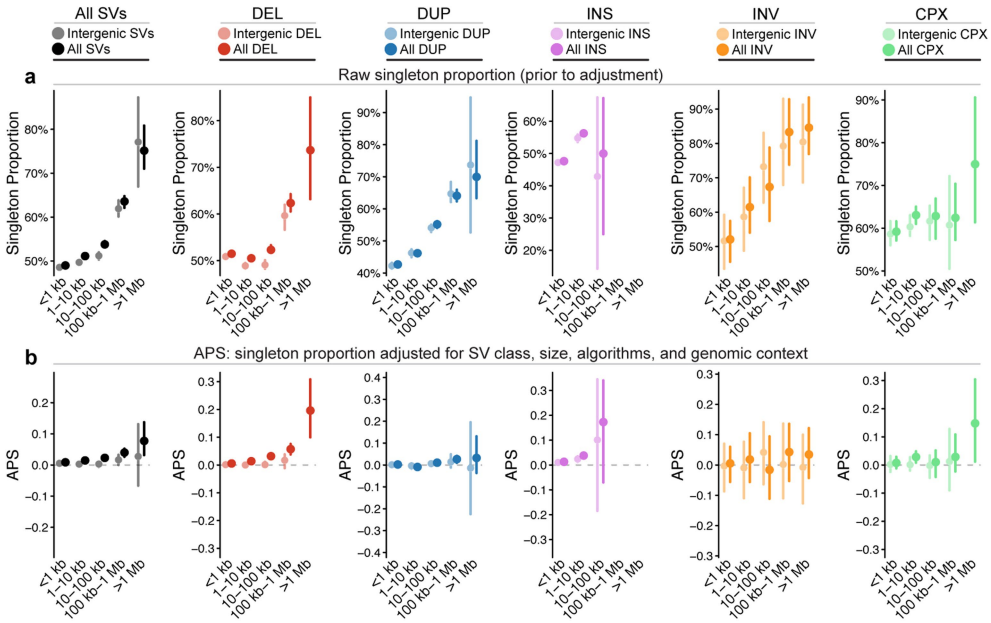
Extended Data Fig. 2 | Benchmarking the technical qualities of the gnomAD-SV callset. We evaluated the quality of gnomAD-SV with seven orthogonal analyses detailed in Supplementary Table 4, Supplementary Figs. 6–9 and Supplementary Note 1. Four core analyses are presented here. **a**, Apparent rates of de novo (that is, spontaneous) heterozygous SVs per child across 970 parent–child trios. Each point is a single trio, and vertical lines denote medians. Given the expected mutation rate of SVs accessible to short-read WGS[1,20] (<1 true de novo SV per trio; see also Fig. 3a), effectively all de novo SVs represented a combination of false-positive genotypes in children and/or false-negative genotypes in parents. SVs passing all filters and included in the final gnomAD-SV callset ('pass') are shown in green. For comparison, variants that did not pass post hoc site-level filters ('not pass') are also shown in purple. **b**, Hardy–Weinberg equilibrium (HWE) metrics for all biallelic SVs localized to autosomes. Deviation from HWE was assessed using a chi-square goodness-of-fit test with one degree of freedom. Vertex labels reflect genotypes: 0/0 denotes homozygous reference; 0/1 denotes heterozygous; and 1/1 denotes homozygous alternate, with all sites shaded by chi-squared *P* value. **c**, Linkage disequilibrium between SVs and SNVs or indels for 23,706 common (allele frequency > 1%) SVs represented as cross-population maximum R^2 values after excluding repetitive and low-complexity regions (see Supplementary Fig. 7). Points and vertical bars represent medians and interquartile ranges, respectively. **d**, Correlation of allele frequency (AF) for 37,907 common SVs captured by both the 1000 Genomes Project and gnomAD-SV[1]. Pearson's correlation coefficient (R^2) is provided.



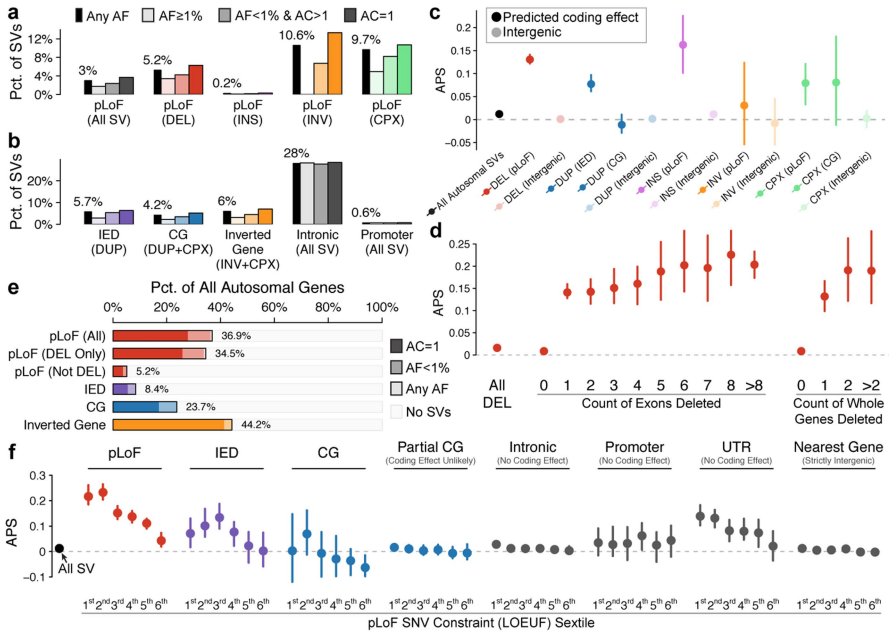
Extended Data Fig. 3 | In silico confirmation of SVs in gnomAD-SV with long-read WGS. We used Pacific Biosciences (PacBio) long-read WGS data available for four samples in this study to perform in silico confirmation to estimate the positive predictive value and breakpoint accuracy for SVs in gnomAD-SV[21,45,46] (Supplementary Fig. 10). **a**, Counts of SVs evaluated per sample in this analysis. SVs were restricted to those with breakpoint-level read support (that is, 'split-read' evidence, 92.8% of all SVs) and did not have breakpoints localized to annotated simple repeats or segmental duplications. **b**, An iterative local long-read WGS realignment algorithm, VaPoR[47], was used to perform in silico confirmation of SVs predicted from short-read WGS in gnomAD-SV. As noted by the VaPoR developers[47], the performance of this approach was sensitive to the sequencing depth of long-read WGS data. Therefore, the weighted mean of the four samples was used as a study-wide long-read WGS confirmation rate, weighting the confirmation rate of each sample based on the square root of its long-read WGS sequencing depth. **c**, Confirmation rates stratified by SV class, size and allele frequency. A mean of 4,829 SVs per sample were assessed. Horizontal green bars denote weighted means.



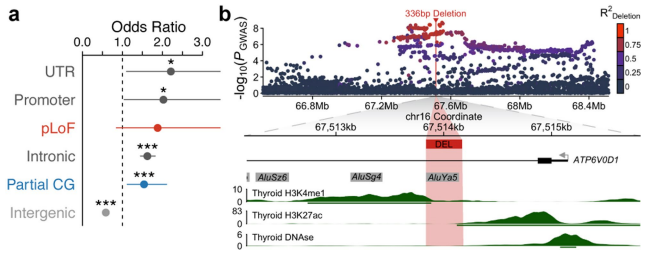
Extended Data Fig. 4 | SVs contribute a substantial burden of rare, homozygous, and coding mutations per genome. a–**d**, Counts of SVs per genome across a variety of parameters, corresponding to median counts of total SVs (**a**), homozygous SVs (**b**), rare SVs (**c**) and singleton SVs (**d**). Samples are grouped by population and coloured by SV types. The solid bar to the left of each population indicates the population median. **e**–**g**, Median counts of genes disrupted by SVs per genome when considering all SVs (including MCNVs) (**e**), homozygous SVs (including MCNVs) (**f**), and rare SVs (**g**). Colours correspond to predicted functional consequence. **h**, Counts of pLoF SVs per genome. For certain categories, such as genes disrupted by rare SVs per genome, a subset of samples (<5%) were enriched above the population average, as expected for individuals carrying large, rare CNVs predicted to cause the disruption of dozens or hundreds of genes (see Extended Data Fig. 1); for the purposes of visualization, the *y* axis for all panels has been restricted to a maximum of three interquartile ranges above the third quartile across all samples for each category.



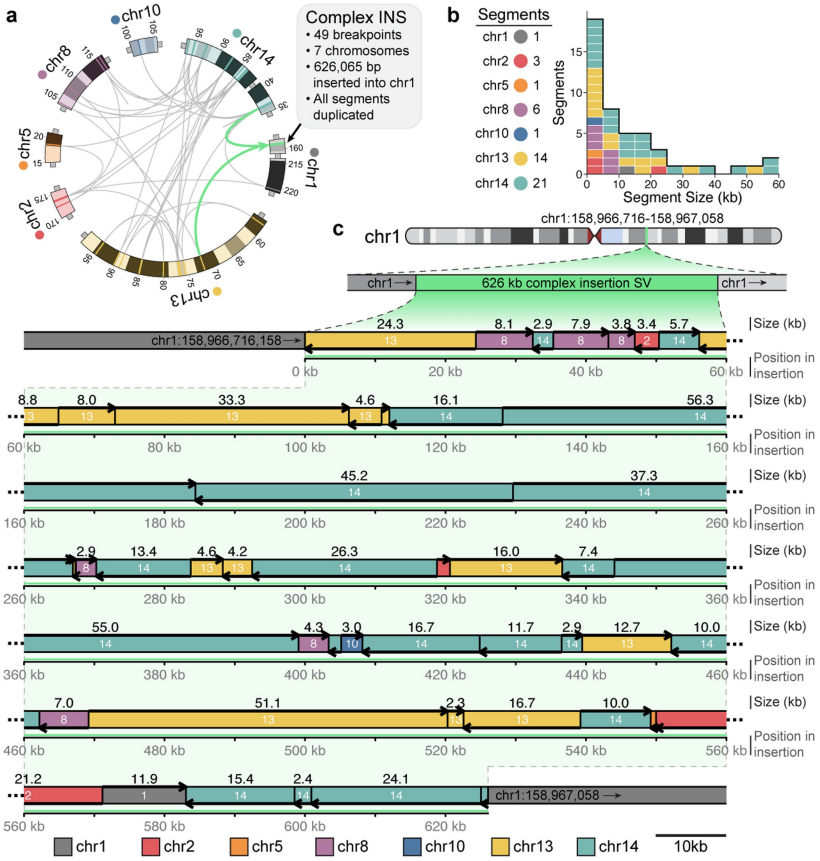
Extended Data Fig. 5 | Rearrangement size is a primary determinant of allele frequency for most classes of SVs. a, Proportion of singleton SVs in five SV size bins for each class of biallelic SVs considered in this study. Intergenic SVs (light colours; $n = 206{,}954$) exhibited reduced singleton proportions when compared to all SVs (dark colours; $n = 335{,}470$) of the same size and class. Bars reflect 95% confidence intervals from 100-fold bootstrapping. Categories with fewer than ten SVs are not shown. **b**, To account for the strong dependency of singleton proportion on SV size and class, we developed the APS metric, which normalizes singleton proportions using SV-specific technical and genomic covariates to permit comparisons of the frequency spectra across SV classes (see Supplementary Fig. 14). The same data as in **a** are shown, transformed onto the APS scale, which shows effectively no dependency on SV size for intergenic SVs. Bars reflect 95% confidence intervals from 100-fold bootstrapping. Residual deviation from APS = 0 is maintained when considering all SVs, owing to APS being intentionally calibrated to intergenic SVs as a proxy for neutral variation. Because larger SVs are more likely to be gene-disruptive, they upwardly bias the APS point estimates due to residual negative selection not captured by SV size alone. Counts of SVs per category for both **a** and **b** are listed in Supplementary Table 9.



Extended Data Fig. 6 | Most SVs within genes appear under negative selection. a, Enrichments for pLoF consequences among rare and singleton SVs across SV classes. **b**, Enrichments for non-pLoF functional consequences among rare and singleton SVs across SV classes. **c**, Adjusted proportion of singletons across SV types and functional consequences. **d**, APS among deletions relative to count of exons and whole genes deleted. **e**, Fractions of all autosomal protein-coding genes with at least one SV across a variety of functional consequences. **f**, Relationship of APS and constraint against pLoF SNVs[4]. For this analysis, intronic, promoter and UTR SVs were required to have precise breakpoints (that is, have 'split-read' support) to protect against any cryptic overlap with coding sequence unable to be annotated due to imprecise breakpoints. For **c**, **d** and **f**, points and vertical bars represent 95% confidence intervals from 100-fold bootstrapping, respectively. Counts of SVs per category in **c** and **d** are provided in Supplementary Table 9. For **d** and **f**, deletions in highly repetitive or low-complexity sequence (≥30% coverage by annotated segmental duplications or simple repeats) were excluded.



Extended Data Fig. 7 | gnomAD-SV can augment disease association studies. a, Functional enrichments of 2,307 common SVs in strong linkage disequilibrium ($R^2 \geq 0.8$) with an SNV associated with a trait or disease in the GWAS catalogue or the UK Biobank[33,34]. Points represent odds ratios of SVs being in strong linkage disequilibrium with at least one GWAS-significant SNV among all SVs in strong linkage disequilibrium with at least one SNV (total $n = 15{,}634$ SVs). Single and triple asterisks correspond to nominal ($P < 0.05$) and Bonferroni-corrected ($P < 0.0083$) significance thresholds from a two-sided Fisher's exact test, respectively. Bars represent 95% confidence intervals. Test statistics, SV counts, and P values are provided in Supplementary Table 6. **b**, Example locus at 16q22.1, where we identified a 336-bp deletion in strong linkage disequilibrium with SNVs significantly associated with hypothyroidism in the UK Biobank[34]. Top, the GWAS signal among genotyped SNVs in the UK Biobank, coloured by strength of linkage disequilibrium (Pearson's R^2 value) with the 336-bp deletion identified in gnomAD-SV. Bottom, the local genomic context of this deletion, which overlaps an annotated intronic *Alu* element near (<1 kb) the first exon of a highly constrained, thyroid-expressed gene, *ATP6V0D1*. The deletion lies amidst histone mark peaks commonly found at active enhancers (H3K27ac and H3K4me1) based on publicly available chromatin data from adult thyroid samples, a phenotype-relevant tissue[48]. Human *Alu* elements are known to frequently act as enhancers, and the sentinel hypothyroidism SNV from the UK Biobank GWAS is a significant expression-modifying variant (that is, eQTL) for *ATP6V0D1* and other nearby genes across many tissues, which indicates that the hypothyroidism risk haplotype modifies expression of *ATP6V0D1* and/or other genes, potentially through the deletion of an intronic enhancer[4,49].



Extended Data Fig. 8 | An extremely complex SV involving 49 breakpoints and seven chromosomes. A highly complex insertion rearrangement from gnomAD-SV in which 47 segments from six different chromosomes were duplicated and inserted into a single locus on chromosome 1, forming a 626,065 bp stretch of contiguous inserted sequence composed of shattered fragments. Given the involvement of multiple chromosomes, the signature of localized shattering, and the clustered breakpoints, we note that this rearrangement has several hallmarks of germline chromothripsis, which has been observed in healthy adults previously, albeit rarely[22]. However, unlike previous reports of germline chromothripsis, there are no apparent whole-chromosome translocations, and all segments were duplicated before being inserted in a compound manner into chromosome 1, potentially suggesting a replication-based repair mechanism. The exact origin of this rearrangement is unclear. **a**, Circos representation of all 49 breakpoints and seven chromosomes involved in this SV. Teal arrows indicate insertion point into chromosome 1. **b**, The median segment size was 8.4 kb. **c**, Linear representation of the rearranged inserted sequence. Colours correspond to chromosome of origin, and arrows indicate strandedness of the inserted sequence, relative to the GRCh37 reference.

nature research

Corresponding author(s): Michael Talkowski, Ph.D.

Last updated by author(s): Feb 27, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All software used for data collection is described in the Methods. All software is publicly available, and all custom software developed in this study has been released under the MIT license. Details on code access are provided in Supplementary Information.
Data analysis	All software used for data analysis is described in the Methods. All software is publicly available, and all custom software developed in this study has been released under the MIT license. Details on code access are provided in Supplementary Information.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The entire gnomAD-SV reference map, from which all conclusions in the study are drawn, has been made publicly accessible through multiple sources. See the Data Availability Statement in the Supplementary Information for more instructions on data access.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical analyses were performed to predetermine sample size. Sample size was dictated by the total number of human whole-genome sequences available with appropriate consent for genetic data aggregation and variant site & frequency release.
Data exclusions	A minority of samples were excluded for various analyses. Where applicable, these exclusions are detailed in the main manuscript, Methods, and Supplementary Information
Replication	No replication was attempted.
Randomization	No randomization was necessary as most analyses were performed on all available samples.
Blinding	No blinding was performed, as no comparisons between subsets of samples were conducted.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The samples analyzed in this study were aggregated across numerous population genetic and complex disease association studies. This aggregation is described in detail in the Methods, and in Karczewski et al., 2019.
Recruitment	Samples were not recruited directly by this study. Data aggregation is described in detail in the Methods, and in Karczewski et al., 2019.
Ethics oversight	This study was overseen by the Broad Institute's Office of Research Subject Protection and the Partners Human Research Committee, and was given a determination of Not Human Subjects Research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Transcript expression-aware annotation improves rare variant interpretation

https://doi.org/10.1038/s41586-020-2329-2

Received: 21 February 2019

Accepted: 23 April 2020

Published online: 27 May 2020

Open access

Check for updates

Beryl B. Cummings[1,2,3], Konrad J. Karczewski[1,2], Jack A. Kosmicki[1,2,4], Eleanor G. Seaby[1,2,5], Nicholas A. Watts[1,2], Moriel Singer-Berk[1], Jonathan M. Mudge[6], Juha Karjalainen[1,2,7], F. Kyle Satterstrom[1,2,7], Anne H. O'Donnell-Luria[1,8,9], Timothy Poterba[1,2,7], Cotton Seed[2,7], Matthew Solomonson[1,2], Jessica Alföldi[1,2], Genome Aggregation Database Production Team*, Genome Aggregation Database Consortium*, Mark J. Daly[1,2] & Daniel G. MacArthur[1,2,145,146] ✉

The acceleration of DNA sequencing in samples from patients and population studies has resulted in extensive catalogues of human genetic variation, but the interpretation of rare genetic variants remains problematic. A notable example of this challenge is the existence of disruptive variants in dosage-sensitive disease genes, even in apparently healthy individuals. Here, by manual curation of putative loss-of-function (pLoF) variants in haploinsufficient disease genes in the Genome Aggregation Database (gnomAD)[1], we show that one explanation for this paradox involves alternative splicing of mRNA, which allows exons of a gene to be expressed at varying levels across different cell types. Currently, no existing annotation tool systematically incorporates information about exon expression into the interpretation of variants. We develop a transcript-level annotation metric known as the 'proportion expressed across transcripts', which quantifies isoform expression for variants. We calculate this metric using 11,706 tissue samples from the Genotype Tissue Expression (GTEx) project[2] and show that it can differentiate between weakly and highly evolutionarily conserved exons, a proxy for functional importance. We demonstrate that expression-based annotation selectively filters 22.8% of falsely annotated pLoF variants found in haploinsufficient disease genes in gnomAD, while removing less than 4% of high-confidence pathogenic variants in the same genes. Finally, we apply our expression filter to the analysis of de novo variants in patients with autism spectrum disorder and intellectual disability or developmental disorders to show that pLoF variants in weakly expressed regions have similar effect sizes to those of synonymous variants, whereas pLoF variants in highly expressed exons are most strongly enriched among cases. Our annotation is fast, flexible and generalizable, making it possible for any variant file to be annotated with any isoform expression dataset, and will be valuable for the genetic diagnosis of rare diseases, the analysis of rare variant burden in complex disorders, and the curation and prioritization of variants in recall-by-genotype studies.

A primary challenge in the use of genome and exome sequencing to predict human phenotypes is that our capacity to identify genetic variation exceeds our ability to interpret their functional impact[3,4]. One underappreciated source of variability for variant interpretation involves differences in alternative mRNA splicing, which enables exons to be expressed at different levels across tissues. These expression differences mean that variants in different regions of a gene can have different phenotypic outcomes depending on the isoforms they affect. For example, variants that occur in an exon differentially included in two isoforms of *CACNA1C* with diverse patterns of tissue expression result in distinct types of Timothy syndrome[5]. Pathogenic variants in the isoform that exhibits multi-tissue expression result in a multi-system disorder[5–7], whereas those on the isoform predominantly expressed in the heart result in more severe and specific cardiac defects[8]. In addition,

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [3]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. [4]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA. [5]Genomic Informatics Group, University Hospital Southampton, Southampton, UK. [6]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [7]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [9]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [145]Present address: Centre for Population Genomics, Garvan Institute of Medical Research, and UNSW Sydney, Syndney, Australia. [146]Present address: Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: d.macarthur@garvan.org.au

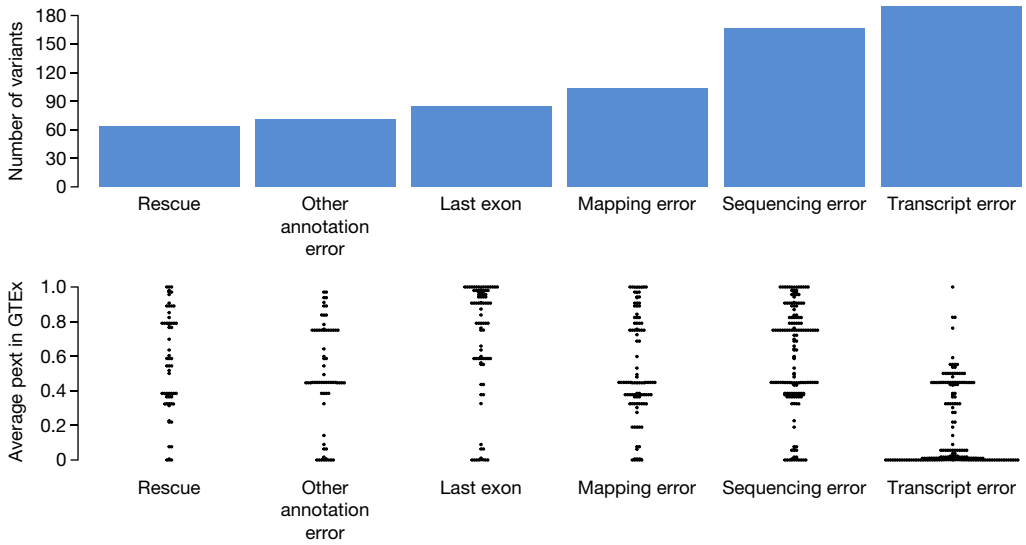


Fig. 1 | Curation of pLoF variants in haploinsufficient disease genes found in gnomAD reveals transcript errors as a major confounding error mode in variant annotation. We identified and manually curated 401 pLoF variants in the gnomAD dataset in 61 haploinsufficient severe developmental delay genes and flagged any reason the pLoF may not be a true LoF variant. Top, the frequency of each error mode present in the 306 variants classified as unlikely to be a true LoF. Transcript errors emerge as a major putative error mode in the annotation of these pLoF variants. Bottom, bee swarm plot shows the average pext score across GTEx tissues for each variant in the error categories. This shows that pext values are discriminately lower for variants that are annotated as possible transcript errors ($P = 4.1 \times 10^{-38}$, two-sided Wilcoxon test between transcript errors and other error modes).

Mendelian disease variants have been found on tissue-specific isoforms[9,10] and isoform expression levels in *TTN* have been used to show that pLoF variants found in healthy controls occur in exons that are absent from dominantly expressed isoforms, whereas those in patients with dilated cardiomyopathy occur on constitutive exons[11], emphasizing the utility of exon expression information for variant interpretation.

Isoform diversity and variant interpretation

We find that isoform diversity is a contributor to the paradoxical finding of disruptive variants in dosage-sensitive disease genes in ostensibly healthy individuals. In the gnomAD database, we identify 401 high-quality pLoF variants that pass both sequencing and annotation quality filters in 61 haploinsufficient disease genes in which heterozygous pLoF variants are established to cause severe developmental delay phenotypes with high penetrance (Methods). Given the severity of these phenotypes and their extremely low prevalence worldwide, ranging from 1 in 10,000 to less than 1 in a million, very few, if any true pLoF variants would be expected to be found in the gnomAD population. As such, most or all of these observed pLoF variants are likely to be sequencing or annotation errors[12]. Manual curation of these variants reveals common error modes that result in probable misannotation of pLoFs, with diversity of transcript structure, mediated by variants falling on low-confidence transcripts, emerging as a major consideration (Fig. 1, Supplementary Fig. 1, Supplementary Tables 1–3). However, no existing tools systematically incorporate information on transcript expression into variant interpretation.

pext score summarizes isoform expression

The advent of large-scale transcriptome sequencing datasets, such as GTEx[2], provides an opportunity to incorporate cross-tissue exon expression into variant interpretation. However, the current formats of these databases do not readily allow for unbiased estimation of exon expression. The GTEx web browser offers information on exon-level read pileup across tissues, but this approach is confounded by technical artefacts such as 3′ bias[13] (preferential coverage of bases close to the 3′ end of a transcript) (Supplementary Fig. 2a). Such systematic biases mean that simple exon-level coverage in a transcriptome dataset cannot be used as a reliable proxy for exon expression, especially in longer genes (Fig. 2a, Supplementary Fig. 2b).

Isoform quantification tools provide estimates of isoform expression levels that correct, albeit imperfectly[13,14], for confounding by 3′ bias as well as other technical artefacts such as isoform length, isoform GC content, and transcript sequence complexity[15–17]. Here, we use isoform-level quantifications from 11,706 tissue samples from the GTEx v7 dataset to derive an annotation-specific expression metric. For each tissue, we annotate each variant with the expression of every possible consequence across all transcripts, which can be used to summarize expression in any combination of tissues of interest. We first compute the median expression of a transcript across tissue samples, and define the expression of a given variant as the sum of the expression of all transcripts for which the variant has the same annotation (Fig. 2a, Supplementary Fig. 3a). By normalizing the expression of the annotation to the total gene expression, we define a metric (proportion expression across transcripts, or 'pext'), which can be interpreted as a measure of the proportion of the total transcriptional output from a gene that would be affected by the variant annotation in question (Supplementary Fig. 3b).

The pext metric allows for quick visualization of the expression of exons across a gene. In Fig. 2b, transcript-expression based annotation is shown for *TCF4*, a haploinsufficient gene in which heterozygous variants result in Pitt–Hopkins syndrome[18], a highly penetrant disorder associated with severe developmental delay. This gene contains 20 unique high-quality pLoF mutations across 56 individuals in the gnomAD database. All 20 variants lie on exons with no evidence of expression across the GTEx dataset (Fig. 2b, Supplementary Fig. 4), which indicates that functional TCF4 protein can be made in the presence of these variants. This visualization is now available for all genes in the gnomAD browser (https://gnomad.broadinstitute.org), and can aid in the rapid identification of variants occurring on exons with little to no evidence of expression in GTEx.

Functional validation of pext

To explore whether expression-based annotation marks functionally important regions, we compared the distribution of the pext metric in

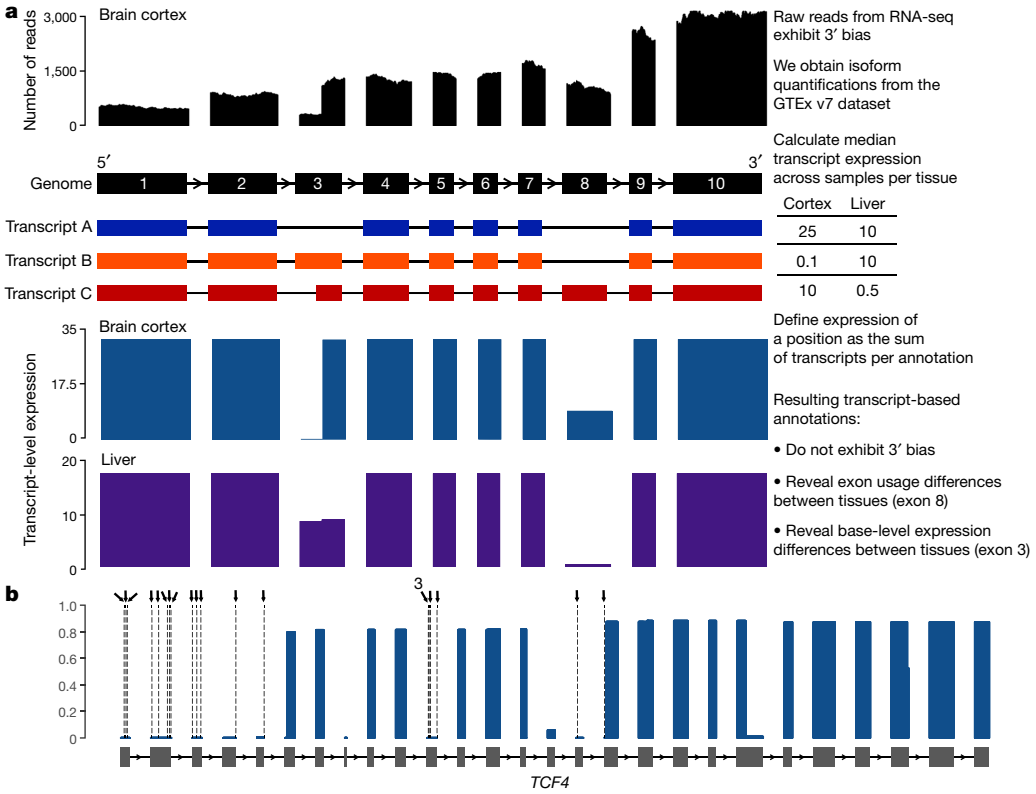


Fig. 2 | Summary of transcript-expression based annotation method. **a**, Overview of transcript-aware annotation. Most genes have many annotated isoforms, which can have varying expression patterns across tissues. Using the number of reads aligning to exonic regions in transcriptome datasets as a proxy for exon expression (top, black) has confounding effects, due to 3′ bias. In this example, although exons 3 and 8 have markedly different expression levels in brain cortex, the number of reads aligning to the two exons is similar, and this masks the differences in exon usage. Transcript-aware annotation defines the expression of every variant as the sum of transcripts that have the same annotation. The resulting transcript-level expression plots do not exhibit 3′ bias, and reveal differences in exon usage, such as those in exons 3 and 8, across tissues. **b**, Example of utility of transcript-expression based annotation. There are 20 high-quality pLoF variants in the haploinsufficient developmental delay gene *TCF4* in gnomAD, annotated as dashed lines and arrows. All 20 variants have no evidence of expression in the GTEx dataset, which suggests that functional TCF4 protein can be made in the presence of these variants.

evolutionarily conserved and unconserved regions using phyloCSF[19]. Exons with patterns of multi-species conservation consistent with coding regions have higher phyloCSF scores, and should exhibit detectable expression patterns, whereas regions with lower scores will be enriched for incorrect exon annotations, which are expected to have little evidence of expression in a population transcriptome dataset. As expected, we observe significantly lower expression for unconserved regions, and near-constitutive expression in highly conserved regions (Fig. 3a, Supplementary Fig. 5a). This difference remains statistically significant after correcting for exon length (logistic regression $P < 1.0 \times 10^{-100}$), which can influence both phyloCSF scores and isoform quantifications, indicating that transcript expression-aware annotation marks functionally relevant exonic regions.

Although the metrics are associated, we find that pext provides orthogonal information to conservation for variant interpretation. For example, regions with low evidence of conservation but high expression (Fig. 3a) are enriched for genes in immune-related pathways (Methods), which are selected for diversity but represent true coding regions. In addition, the pext value is higher for pLoF variants annotated as high confidence by the loss-of-function transcript effect estimator (LOFTEE) package[1], with no additional flags than those flagged as having found on unlikely open-reading frames or weakly conserved regions (Fig. 3b, Supplementary Fig. 5b). However, high-confidence LOFTEE variants with no flags can also have low pext values, which suggests that transcript-expression-aware annotation adds additional information to the currently available interpretation toolkit.

We undertook manual evaluation of 128 regions marked as unexpressed (mean pext < 0.1 in all tissues and in GTEx brain) in 61 haploinsufficient genes following the GENCODE manual annotation workflow[20] to evaluate the annotation quality in these coding sequence (CDS) regions. One-third of flagged regions were associated with low-quality models that have been removed or switched to non-coding biotypes in subsequent GENCODE releases (Supplementary Fig. 6), and 70% of the remaining regions correspond to models that satisfy only minimum criteria for inclusion in the gene set, corresponding to 'putative' annotations that lack markers for CDS functionality (Supplementary Table 4). Nonetheless, we find support for some highly conserved CDS regions, several of which show evidence of transcription in fetal tissues, underlining the importance of incorporating several isoform expression datasets for interpretation (Supplementary Fig. 6d).

Non-synonymous variants found on constitutively expressed regions would be expected to be more deleterious than those on regions with no evidence of expression. To test this, we defined expression bins based on the average pext value across GTEx tissues, in which an average pext value less than 0.1 was defined as low (or unexpressed), above 0.9 as high (or near-constitutive) and intermediate values as medium expression. We compared the mutability-adjusted proportion singleton (MAPS), a measure of negative selection on variant classes[21], partitioned on the loss-of-function observed/expected upper-bound fraction (LOEUF) decile, a measure of constraint against pLoF variants in the gnomAD dataset[1] in each of these expression bins. MAPS scores differed substantially between pLoF variants found on low-expressed and high-expressed regions in genes intolerant to pLoF variation (Fig. 3c,

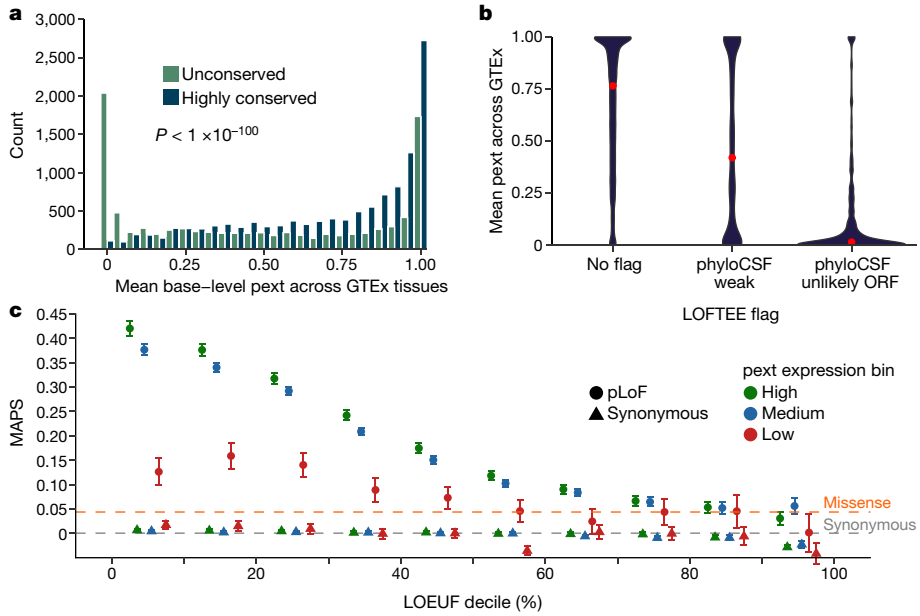


Fig. 3 | Functional validation of transcript-expression based annotation. **a**, We define highly conserved and unconserved regions as phyloCSF > 1,000 (n = 9,817) and phyloCSF < −100 (n = 11,860), respectively, and compare the expression status of these regions across GTEx. Regions with high phyloCSF scores are enriched for near-constitutive expression, whereas unconserved regions are enriched for little to no usage across GTEx. This difference is significant after correcting for gene length (logistic regression $P < 1 \times 10^{-100}$). We note that unconserved regions with high levels of expression (pext > 0.9) are enriched for immune-related genes, which are selected for diversity and thus have low conservation, but represent true coding regions. **b**, Transcript-expression based annotation recapitulates, and adds information to, existing interpretation tools. High-confidence pLoF LOFTEE variants in gnomAD with no flags (n = 458,880) are enriched for higher pext values, whereas high-confidence pLoF variants falling on low phyloCSF (n = 44,373) or unlikely open-reading frame regions (n = 2,437) are enriched for low expression. However, high-confidence pLoF variants can also have a low pext score. Variants flagged falling on regions that are unlikely open-reading frame or have weak conservation are enriched for lower pext values. Red dots denote the median pext value across GTEx, **c**, Non-synonymous variants found on near-constitutive regions tend to be more deleterious. We compared the MAPS score for variants with low (<0.1), medium (0.1 ≤ pext ≤ 0.9) and high (pext > 0.9) expression. Variants with near-constitutive expression have a higher MAPS score, which indicates higher deleteriousness than those with little to no evidence of expression. Points represent MAPS values and error bars denote the 95% confidence interval. Dashed grey and orange lines represent MAPS values for all gnomAD missense and synonymous variants, respectively. The number of variants evaluated per category and unadjusted proportion singleton values can be found in Supplementary Table 5a.

Supplementary Fig. 5c, Supplementary Table 5a, b). This information is complementary to existing variant prioritization tools such as PolyPhen-2[22] (Supplementary Fig. 5d, Supplementary Table 5c). This skew of non-synonymous variation in high-expressed regions suggests that variation arising in such exons tends be more deleterious, whereas non-synonymous variants on regions with low expression are similar to missense variants in their inferred deleteriousness.

Use of pext in variant interpretation

To evaluate the utility of transcript expression-based annotation in Mendelian variant interpretation, we assessed the number of variants that would be filtered based on a pext cut-off value of less than 0.1 (low expression) across GTEx tissues for three gene sets. First, we evaluated high-quality pLoF variants in the 61 manually curated haploinsufficient genes in gnomAD and ClinVar[23]. The low pext expression bin resulted in filtering of 22.8% of pLoF variants in haploinsufficient developmental delay genes in gnomAD, but only 3.8% of high-quality pathogenic variants in ClinVar ($P = 4.7 \times 10^{-35}$) (Fig. 4a, Methods). We next compared pLoF variants in autosomal recessive disease genes found in a homozygous state in at least one individual in gnomAD and any pLoF variant in these genes in ClinVar and observed similar results: expression-based annotation filters 30.0% of variants in gnomAD while only filtering 3.2% of variants in ClinVar (Fig. 4b) ($P = 3.5 \times 10^{-61}$).

Finally, we evaluated gnomAD pLoF variants in genes that are constrained against pLoF variation[1] (LOEUF score < 0.35). Given that these genes are depleted for loss-of-function variation in the general population, we expect the observed pLoF variants in these genes to be enriched for annotation errors. We compared the proportion filtered to synonymous variants in the same genes, which we expect to be randomly distributed. Our metric removes 16.8% of pLoF variants in constrained genes, but only 5.2% of synonymous variants (Fig. 4c) ($P < 1.0 \times 10^{-100}$). In all cases, the vast majority of filtered variants were otherwise high-confidence with no LOFTEE annotation flags, which suggests again that pext provided additional information to existing variant prioritization tools in removing annotation errors (Supplementary Fig. 7).

Use of pext in burden testing

To explore the benefits of this approach for rare variant analysis, we applied pext binning to burden testing of de novo variants in patients with developmental delay/intellectual disability (DD/ID) or autism spectrum disorder (ASD) using a set of 23,970 de novo variants collated from several studies including the Deciphering Developmental Disorders (DDD) project and the Autism Sequencing Consortium (ASC)[24–29]. We find that de novo pLoF variants in patients with DD/ID in low-expressed regions have similar effect sizes to those of synonymous variants (rate ratio of low-expressed pLoFs = 1.08, P = 0.90), whereas pLoF variants in highly expressed regions have much larger effect sizes (rate ratio = 4.64, $P = 3.74 \times 10^{-38}$) (Fig. 5a). This observation is consistent for de novo variants in autism (rate ratio for low-expressed pLoFs = 0.80, P = 0.47; rate ratio for high-expressed pLoFs = 2.11, $P = 8.2 \times 10^{-8}$) (Fig. 5b) and congenital heart disease with co-morbid neurodevelopmental delay (Supplementary Fig. 8a) as well as rare variants (allele count ≤ 10) identified in highly constrained genes in the large iPSYCH case–control study

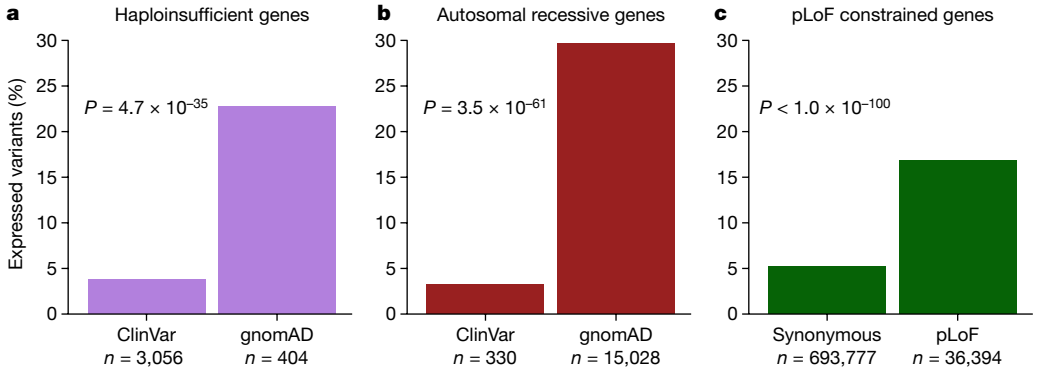


Fig. 4 | Transcript-expression based annotation aids Mendelian variant interpretation. a, Comparison of the proportion of high-quality pLoF variants filtered in a curated list of 61 haploinsufficient developmental delay genes in gnomAD versus ClinVar with a cut-off value of average pext across GTEx ≤ 0.1 (low expression). Expression-based filtering results in removal of 22.8% of gnomAD pLoFs and 3.8% of confidently curated set of pLoFs in ClinVar. **b**, Expression-based annotation filters 30% of pLoF variants found in gnomAD in a homozygous state in at least one individual, and 3.2% of any pLoF variants found in the same genes in ClinVar. **c**, We extended this filtering approach to pLoF and synonymous variants in gnomAD pLoF-intolerant genes (defined by LOEUF < 0.35). This filters 16.8% of LoF and 5.2% of synonymous variants. The total number of high-quality variants considered in each group is shown. For all pLoFs only high-confidence LOFTEE variants were considered. *P* values were determined by two-sided Fisher's exact test for counts.

of Danish patients with autism spectrum disorder and attention-deficit/hyperactivity disorder (Supplementary Fig. 8b). Overall, we consistently observe low-expressed pLoFs to have effect sizes similar to those of synonymous variants, with pLoF variants in constitutive regions having larger effect sizes, which suggests that incorporating transcript expression-aware annotation in rare variant studies can boost power for gene discovery.

Discussion

We have described the development and validation of a transcript expression-based annotation framework to integrate results from transcriptome sequencing experiments into clinical variant interpretation. Although our initial analysis uses GTEx, our method can be used with any isoform expression dataset to annotate any variant file rapidly in the scalable software framework Hail (https://hail.is). For example, annotation of more than 120,000 gnomAD individuals with GTEx takes under an hour using 60 cores, at a cost of about US$5 on public cloud compute, which can be further scaled to larger datasets. In addition, the annotations we provide are flexible: although we have described the use of average transcript-level expression across many tissues, alternative approaches such as using maximum expression across any tissue may prove useful depending on variant interpretation goals (Supplementary Figs. 9, 10).

We note that although this metric successfully discriminates between near-constitutive and low expression levels, which are useful for prioritizing and filtering variants, respectively, regions with intermediate expression levels are more challenging to interpret. However, we hypothesize directed analyses of intermediate expression levels may help to determine the role of alternative splicing in phenotypic diversity[30,31]. In addition, although we have binned average pext scores across GTEx tissues into low, medium and high expression, different genes will probably have varying optimal tissues and thresholds for variant interpretation. Regions tagged as low expression are often corroborated by expert opinion of CDS curation, but domain knowledge of a gene will outperform this summary metric.

An important caveat in our approach is the imprecision of isoform quantification methods using short-read transcriptome data. However, we note that repeating key analyses in the manuscript with a different isoform quantification tool showed consistent results (Methods, Supplementary Fig. 11, Supplementary Table 6), suggesting robustness to the precise pipeline used. The utility of this framework will increase

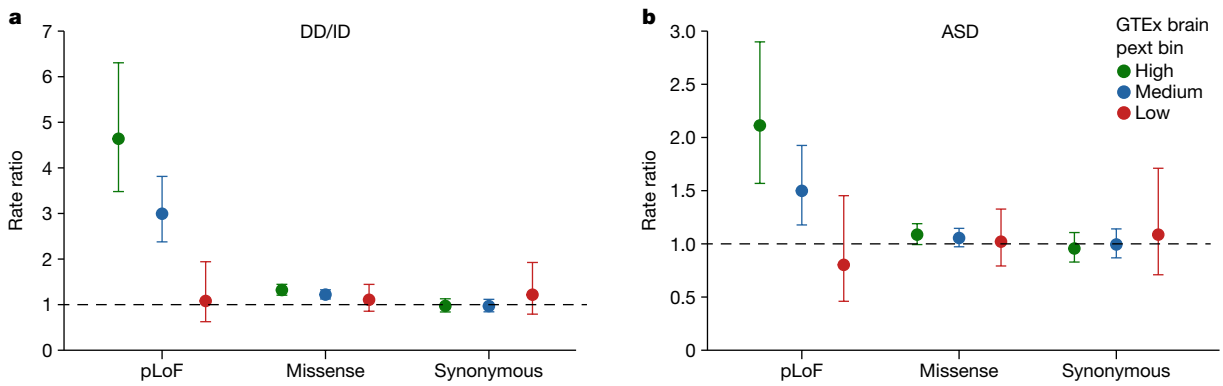


Fig. 5 | Application of transcript-expression based annotation to de novo variant analyses in ASD and DD/ID. a, **b**, Transcript-expression-based analyses in patients with DD/ID (**a**) or ASD (**b**). We find that de novo pLoF variants found on near-constitutively expressed regions in GTEx brain tissues have larger effect sizes than de novo LoF variants in weakly expressed regions in both disorders. Notably, de novo pLoF variants found on regions with little evidence for expression are as equally distributed in cases versus controls as de novo synonymous variants, which suggests that such variants can be removed from analyses of gene burden testing to boost discovery power. The high pext expression bin contains 46.1%, 42.3% and 11.4%, and the low-expression bin contains 4.0%, 6.0% and 11.4% of 1,249, 752 and 166 de novo pLoF variants found in patients with DD/ID, ASD and controls, respectively. Points represent rate ratio estimate and error bars represent 95% confidence interval from the Poisson exact test.

as our ability to quantify isoform expression across tissues improves, including refinement of methods and gene models, as well as availability of long-read RNA-sequencing data from human tissues. In addition, the improvement of single-cell RNA-seq technologies and the generation of data across human tissues will provide insight into cell type-specific exon usage for incorporation into variant interpretation[32].

The code used to generate pext is available as open source software (https://github.com/macarthur-lab/tx_annotation). In addition, we provide a precomputed file of the transcript expression value for every possible single nucleotide variant in the human genome. This metric has already proven useful in variant curation for the identification of drug targets[33] and for filtering variants for the identification of human knockouts[1]. Overall, our metric can be incorporated into variant interpretation in Mendelian disease pipelines, analyses of rare variant burden, and the prioritization of variants for recall-by-genotype studies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2329-2.

1. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* https://doi.org/10.1038/s41586-020-2308-7 (2020).
2. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
3. MacArthur, D. G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
4. Goldstein, D. B. et al. Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* **14**, 460–470 (2013).
5. Dick, I. E., Joshi-Mukherjee, R., Yang, W. & Yue, D. T. Arrhythmogenesis in Timothy Syndrome is associated with defects in Ca^{2+}-dependent inactivation. *Nat. Commun.* **7**, 10370 (2016).
6. Splawski, I. et al. $Ca_V1.2$ calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* **119**, 19–31 (2004).
7. Liao, P. & Soong, T. W. CaV1.2 channelopathies: from arrhythmias to autism, bipolar disorder, and immunodeficiency. *Pflugers Arch.* **460**, 353–359 (2010).
8. Splawski, I. et al. Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. *Proc. Natl Acad. Sci. USA* **102**, 8089–8096 (2005).
9. Nousbeck, J. et al. A mutation in a skin-specific isoform of SMARCAD1 causes autosomal-dominant adermatoglyphia. *Am. J. Hum. Genet.* **89**, 302–307 (2011).
10. Guven, A. & Tolun, A. TBC1D24 truncating mutation resulting in severe neurodegeneration. *J. Med. Genet.* **50**, 199–202 (2013).
11. Roberts, A. M. et al. Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Sci. Transl. Med.* **7**, 270ra276 (2015).
12. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
13. Zheng, W., Chung, L. M. & Zhao, H. J. B. B. Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics* **12**, 290 (2011).
14. Teng, M. et al. A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 74 (2016).
15. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
16. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
17. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
18. Sweatt, J. D. Pitt-Hopkins Syndrome: intellectual disability due to loss of TCF4-regulated gene transcription. *Mol. Med.* **45**, e21 (2013).
19. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
20. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47** (D1), D766–D773 (2019).
21. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
22. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Prot. Human Genet.* **Chapter 7**, Unit7.20–Unit27.20 (2013).
23. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46** (D1), D1062–D1067 (2018).
24. Deciphering Developmental Disorders Study. Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
25. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584. e23 (2020).
26. Lelieveld, S. H. et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).
27. Hamdan, F. F. et al. De novo mutations in moderate or severe intellectual disability. *PLoS Genet.* **10**, e1004772–e1004772 (2014).
28. de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
29. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
30. Tress, M. L., Abascal, F. & Valencia, A. Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.* **42**, 98–110 (2017).
31. Blencowe, B. J. The relationship between alternative splicing and proteomic complexity. *Trends Biochem. Sci.* **42**, 407–408 (2017).
32. Regev, A. et al. The Human Cell Atlas. *eLife*. **6**, e27041 (2017).
33. Minikel, E. V. et al. Evaluating drug targets through human loss-of-function genetic variation. *Nature* https://doi.org/10.1038/s41586-020-2267-z (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Genome Aggregation Database Production Team

Jessica Alföldi[1,2], Irina M. Armean[1,2,6], Eric Banks[10], Louis Bergelson[10], Kristian Cibulskis[10], Ryan L. Collins[1,4,11], Kristen M. Connolly[12], Miguel Covarrubias[10], Beryl B. Cummings[1,2,3], Mark J. Daly[1,2], Stacey Donnelly[1], Yossi Farjoun[10], Steven Ferriera[13], Laurent Francioli[1,2], Stacey Gabriel[13], Laura D. Gauthier[10], Jeff Gentry[10], Namrata Gupta[1,13], Thibault Jeandet[10], Diane Kaplan[10], Konrad J. Karczewski[1,2], Kristen M. Laricchia[1,2], Christopher Llanwarne[10], Eric V. Minikel[1,2,7], Ruchi Munshi[10], Benjamin M. Neale[1,2], Sam Novod[10], Anne H. O'Donnell-Luria[1,8,9], Nikelle Petrillo[10], Timothy Poterba[1,2,7], David Roazen[10], Valentin Ruano-Rubio[10], Andrea Saltzman[1], Kaitlin E. Samocha[14], Molly Schleicher[1], Cotton Seed[2,7], Matthew Solomonson[1,2], Jose Soto[10], Grace Tiao[1,2], Kathleen Tibbetts[10], Charlotte Tolonen[10], Christopher Vittal[2,7], Gordon Wade[10], Arcturus Wang[1,2,7], Qingbo Wang[1,2,4], James S. Ware[1,15,16], Nicholas A. Watts[1,2], Ben Weisburd[10] & Nicola Whiffin[1,15,16]

[10]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [11]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [12]Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [13]Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [14]Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK. [15]National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. [16]Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London, UK.

Genome Aggregation Database Consortium

Carlos A. Aguilar Salinas[17], Tariq Ahmad[18], Christine M. Albert[19,20], Diego Ardissino[21], Gil Atzmon[22,23,24], John Barnard[25], Laurent Beaugerie[26], Emelia J. Benjamin[27,28,29], Michael Boehnke[30], Lori L. Bonnycastle[31], Erwin P. Bottinger[32], Donald W. Bowden[33,34,35], Matthew J. Bown[36,37], John C. Chambers[38,39,40], Juliana C. Chan[41], Daniel Chasman[19], Judy Cho[32], Mina K. Chung[25], Bruce Cohen[42,43], Adolfo Correa[44], Dana Dabelea[45], Mark J. Daly[1,2], Dawood Darbar[46], Ravindranath Duggirala[47], Josée Dupuis[27,48], Patrick T. Ellinor[1,49], Roberto Elosua[50,51,52], Jeanette Erdmann[53,54,55], Tõnu Esko[1,56], Martti Färkkilä[57], Jose Florez[1], Andre Franke[58], Gad Getz[59], Benjamin Glaser[60], Stephen J. Glatt[61], David Goldstein[62,63], Clicerio Gonzalez[64], Leif Groop[65], Christopher Haiman[66], Craig Hanis[67], Matthew Harms[68,69], Mikko Hiltunen[70], Matti M. Holi[71], Christina M. Hultman[72], Mikko Kallela[73], Jaakko Kaprio[74,75], Sekar Kathiresan[1,11,76], Bong-Jo Kim[77], Young Jin Kim[77], George Kirov[78], Jaspal Kooner[39,40,79], Seppo Koskinen[80], Harlan M. Krumholz[81], Subra Kugathasan[82], Soo Heon Kwak[83], Markku Laakso[84,85], Terho Lehtimäki[86], Ruth J. F. Loos[32,87], Steven A. Lubitz[1,88], Ronald C. W. Ma[89,90,91], Daniel G. MacArthur[1,2], Jaume Marrugat[51,92], Kari M. Mattila[86], Steven McCarroll[7,93], Mark I. McCarthy[94,95,96], Dermot McGovern[97], Ruth McPherson[98], James B. Meigs[1,76,99], Olle Melander[100], Andres Metspalu[56], Benjamin M. Neale[1,2], Peter M. Nilsson[101], Michael C. O'Donovan[78], Dost Ongur[42,76], Lorena Orozco[102], Michael J. Owen[78], Colin N. A. Palmer[103], Aarno Palotie[1,2,74], Kyong Soo Park[83,104], Carlos Pato[105], Ann E. Pulver[106], Nazneen Rahman[107], Anne M. Remes[108], John D. Rioux[109,110], Samuli Ripatti[1,74,111], Dan M. Roden[112,113], Danish Saleheen[114,115,116], Veikko Salomaa[117], Nilesh J. Samani[36,37], Jeremiah Scharf[1,7,11], Heribert Schunkert[118,119], Moore B. Shoemaker[120], Pamela Sklar[121,122,123,147], Hilkka Soininen[124],

Harry Sokol[26], Tim Spector[125], Patrick F. Sullivan[72,126], Jaana Suvisaari[117], E. Shyong Tai[127,128,129], Yik Ying Teo[127,130,131], Tuomi Tiinamaija[74,132,133], Ming Tsuang[134,135], Dan Turner[136], Teresa Tusie-Luna[137,138], Erkki Vartiainen[139], James S. Ware[1,15,16], Hugh Watkins[140], Rinse K. Weersma[141], Maija Wessman[74,132], James G. Wilson[142] & Ramnik J. Xavier[143,144]

[17]Unidad de Investigacion de Enfermedades Metabolicas, Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico. [18]Peninsula College of Medicine and Dentistry, Exeter, UK. [19]Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [20]Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [21]Department of Cardiology, University Hospital, Parma, Italy. [22]Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel. [23]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, USA. [24]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. [25]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [26]Sorbonne Université, APHP, Gastroenterology Department, Saint Antoine Hospital, Paris, France. [27]Framingham Heart Study, National Heart, Lung, & Blood Institute and Boston University, Framingham, MA, USA. [28]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [29]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA. [30]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [31]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [32]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [33]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [34]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [35]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [36]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK. [37]NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK. [38]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [39]Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK. [40]Imperial College Healthcare NHS Trust, Imperial College London, London, UK. [41]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, Hong Kong. [42]Program for Neuropsychiatric Research, McLean Hospital, Belmont, MA, USA. [43]Department of Psychiatry, Harvard Medical School, Boston, MA, USA. [44]Department of Medicine, University of Mississippi Medical Center, Jackson, MI, USA. [45]Department of Epidemiology, Colorado School of Public Health, Aurora, CP, USA. [46]Department of Medicine and Pharmacology, University of Illinois at Chicago, Chicago, IL, USA. [47]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA. [48]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [49]Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [50]Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain. [51]Centro de Investigación Biomédica en Red Enfermedades Cardiovaculares (CIBERCV), Barcelona, Spain. [52]Department of Medicine, Medical School, University of Vic-Central University of Catalonia, Vic, Spain. [53]Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany. [54]DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, Lübeck, Germany. [55]University Heart Center Lübeck, Lübeck, Germany. [56]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [57]Clinic of Gastroenterology, Helsinki University and Helsinki University Hospital, Helsinki, Finland. [58]Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany. [59]Cancer Genome Computational Analysis Group, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [60]Endocrinology and Metabolism Department, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [61]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA. [62]Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [63]Department of Genetics & Development, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [64]Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Cuernavaca, Mexico. [65]Genomics, Diabetes and Endocrinology, Lund University, Lund, Sweden. [66]Lund University Diabetes Centre, Malmö, Sweden. [67]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA. [68]Department of Neurology, Columbia University, New York, NY, USA. [69]Institute of Genomic Medicine, Columbia University, New York, NY, USA. [70]Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland. [71]Department of Psychiatry, Helsinki University Central Hospital, Lapinlahdentie, Helsinki, Finland. [72]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [73]Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland. [74]Institute for Molecular Medicine FIMM, University of Helsinki, Helsinki, Finland. [75]Department of Public Health, University of Helsinki, Helsinki, Finland. [76]Department of Medicine, Harvard Medical School, Boston, MA, USA. [77]Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, South Korea. [78]MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Cardiff, UK. [79]National Heart and Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. [80]Department of Health, THL-National Institute for Health and Welfare, Helsinki, Finland. [81]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA. [82]Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, Georgia, USA. [83]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. [84]The University of Eastern Finland, Institute of Clinical Medicine, Kuopio, Finland. [85]Kuopio University Hospital, Kuopio, Finland. [86]Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [87]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [88]Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA, USA. [89]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. [90]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [91]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. [92]Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain. [93]Department of Genetics, Harvard Medical School, Boston, MA, USA. [94]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. [95]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [96]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [97]F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [98]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada. [99]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [100]Department of Clinical Sciences, University Hospital Malmo Clinical Research Center, Lund University, Malmo, Sweden. [101]Lund University, Dept. Clinical Sciences, Skane University Hospital, Malmo, Sweden. [102]Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, Mexico. [103]Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. [104]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea. [105]Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, CA, USA. [106]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [107]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK. [108]Research Unit of Clinical Neuroscience, University of Oulu, Oulu, Finland. [109]Research Center, Montreal Heart Institute, Montreal, Quebec, Canada. [110]Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada. [111]Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [112]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [113]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [114]Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [115]Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [116]Center for Non-Communicable Diseases, Karachi, Pakistan. [117]National Institute for Health and Welfare, Helsinki, Finland. [118]Deutsches Herzzentrum München, Munich, Germany. [119]Technische Universität München, Munich, Germany. [120]Division of Cardiovascular Medicine, Nashville VA Medical Center and Vanderbilt University, School of Medicine, Nashville, TN, USA. [121]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [122]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [123]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [124]Institute of Clinical Medicine Neurology, University of Eastern Finland, Kuopio, Finland. [125]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [126]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [127]Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore, Singapore. [128]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [129]Duke-NUS Graduate Medical School, Singapore, Singapore. [130]Life Sciences Institute, National University of Singapore, Singapore, Singapore. [131]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore. [132]Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland. [133]HUCH Abdominal Center, Helsinki University Hospital, Helsinki, Finland. [134]Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego, CA, USA. [135]Institute of Genomic Medicine, University of California, San Diego, CA, USA. [136]Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel. [137]Instituto de Investigaciones Biomédicas UNAM, Mexico City, Mexico. [138]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. [139]Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland. [140]Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [141]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. [142]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [143]Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [144]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. [147]Deceased: Pamela Sklar.

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Curation of pLoF variants in haploinsufficient developmental disease genes

To identify haploinsufficient developmental delay genes, we selected genes curated by the ClinGen Dosage Sensitivity Working Group[34] 58 of the 61 genes had a score of 3 with sufficient evidence for pathogenicity, whereas two genes (*CHAMP1*, *CTCF*) had a score of 2 (some evidence) and one gene (*RERE*) was not yet scored. The penetrance of pathogenic variants in each gene was reviewed in the literature, and only genes with more than 75% reported penetrance were included. These conditions are those too severe to expect to see an individual in gnomAD (probably unable to consent for a study without guardianship). The 61 genes include 50 autosomal genes of high severity and high penetrance and 11 genes on chromosome X in which the phenotype is expected to be severe or lethal in males and moderate to severe in females. The resulting gene list is available at gs://gnomad-public/papers/2019-tx-annotation/data/gene_lists/HI_genes_100417.tsv.

We extracted pLoF variants, defined as essential splice acceptor, essential splice donor, stop-gained, and frameshift variants, identified in the 61 haploinsufficient disease genes from the gnomAD v2.1.1 exome and genome sites tables, and considered only those pLoF variants that passed random forest filtering in the gnomAD dataset, and were annotated as high confidence by LOFTEE v1.0. Of 61 genes, 55 had at least one high-quality pLoF available in gnomAD. We performed manual curation of 401 pLoF variants using a web-based curation portal to identify any reason a pLoF may have been a variant calling or annotation error, and categorized the likelihood of each variant being a true LoF.

Evidence for classifying an LoF variant as artefactual was categorized into the following groups: mapping error, strand bias, reference error, genotyping error, homopolymer sequence, in-frame multi-nucleotide variant or frame-restoring indel, essential splice site rescue, minority of transcripts, weak exon conservation, last exon, and other annotation error. All possible reasons also to reject a LoF consequence were flagged, even when a single criterion would categorize the variant as not LoF. Variants were then categorized as LoF, probable LoF, probably not LoF, and not LoF based on criteria outlined in Supplementary Table 2. Supplementary Fig. 1a shows the distribution of the LoF verdicts for the 401 pLoF variants.

Technical errors comprised genotyping errors, strand biases, reference errors, and repetitive regions that could be detected by visual inspection of reads in the Integrative Genomics Viewer[35] (IGV) and from the UCSC genome browser[36]. Genotyping errors comprised skewed allele balances (conservative cutoff of ≤ 35%), low complexity sequences, GC-rich regions, homopolymer tracts (≥6 base pairs or ≥ 6 trinucleotide repeats) and low quality metrics (genotype quality < 20). Strand bias was flagged when a variant was skewed preferentially on the forward or reverse strand, or when the majority (>90%) of a given strand covered a region; this was often observed around intron–exon boundaries. Strand biases despite balanced coverage of the forward and reverse strands were weighted towards probably not LoF, whereas a strand bias due to skewed strand coverage was weighted alongside other genotyping errors. Reference errors were uncommon, but identified by a small deletion in a given exon, posing as a <5-base-pair intron. Most genotyping errors and strand biases in isolation were not deemed critical in deciding whether a variant was probably not LoF or not LoF, with the exception of allele balance ≤25%. Mapping errors were often identified by an enrichment of complex variation surrounding a variant of interest. Furthermore, the UCSC browser was used to highlight mapping discrepancies, such as self-chain alignments, segmental duplications, simple tandem repeats, and microsatellite regions.

In-frame multi-nucleotide variants (MNVs), essential splice site rescue, and frame-restoring insertion-deletions are rescue events that are predicted to restore gene function. MNVs were visualized in IGV and cross checked with codons from the UCSC browser; in frame MNVs that rescued stop codons were scored as not LoF. Essential splice site rescue occurs when an in frame alternative donor or acceptor site is present, which probably has a minimal effect on the transcript. A total of 36 base pairs upstream and downstream of the splice variant were assessed for splice site rescue. Cryptic splice sites within 6 base pairs of the splice variant were considered a complete rescue, rendering the variant not LoF. Rescue sites >6 base pairs away but within ±20 base pairs were weighted with less confidence, scoring as probably not LoF. All potential splice site rescues were validated using Alamut v.2.11 (https://www.interactive-biosoftware.com/alamut-visual/). Frame-restoring indels were identified by scanning approximately ±80 base pairs from the annotated indel and counting any insertions/deletions to assess if the frame would be restored.

Transcript errors encompass issues surrounding alternative transcripts, variants within a terminal coding exon, poorly conserved exons, and re-initiation events. Coding variants that occupied the minority (<50%) of NCBI coding RefSeq transcripts for a given gene were considered not LoF. These variants often affected poorly conserved exons, as determined by PhyloP[37], PhyloCSF[19] and visualization in the UCSC browser[36]. The only exceptions to the minority of transcript criteria were cases where the exon was well conserved, which relegated the categorization to probably not LoF. Variants within the last coding exon, or within 50 base pairs of the penultimate coding exon were also considered not LoF, unless $25\% < x < 50\%$ of the coding sequence was affected, in which case the variant was deemed probably not LoF. If >50% of the coding sequence was disrupted by a variant in the last exon, this was deemed probably LoF. Other transcript errors included: re-initiation errors; upstream stop codons of a given LoF variant; variants that fell on exactly 50% of coding RefSeq transcripts; and/or partial exon conservation. Re-initiation events were flagged when a methionine downstream of the variant in the first coding exon was predicted to restart transcription, and were predicted to be probably not LoF. Variants occurring after a stop codon in the last coding exon were considered not LoF, particularly across the region of the exon or transcript in question. Error categories were grouped for Fig. 1 as follows: Minority of transcripts and weak exon conservation were grouped as transcript errors, genotyping errors and homopolymers as sequencing errors, essential splice rescue and MNV grouped as rescue and strand bias was included in other annotation errors.

The criteria above were strictly adhered throughout and manual curation was performed by two independent reviewers to ensure maximum consistency and minimize human error. Any discordance in curation was re-curated by both curators together and resolved. Full results of manual curation are available in Supplementary Table 3.

Calculation of transcript-expression aware annotation

We first imported the GTEx v7 isoform quantifications into Hail and calculated the median expression of every transcript per tissue. This precomputed summary isoform expression matrix is available for GTEx v7 in gs://gnomad-public/papers/2019-tx-annotation/data/GRCH37_hg19/. We also import and annotate a variant file with the Variant Effect Predictor (VEP) version 85[38] against Gencode v19[20], implemented in Hail with the LOFTEE v1.0 plugin.

We use the transcript consequences VEP field to calculate the sum of isoform expression for variant annotations, that is, the annotation-level expression across transcripts (ext). For variants that have multiple consequences for one transcript (for example, a single nucleotide variant

that is both a missense and a splice region variant on one transcript) we use the worst consequence, ordered by VEP (in this example, missense takes precedence over splice region). We filter the consequences to those only occurring on protein coding transcripts. Full ordering of the VEP consequences is available at: useast.ensembl.org/info/genome/variation/prediction/predicted_data.html

We then sum the expression of every transcript per variant, for every combination of consequence, LOFTEE filter, and LOFTEE flag for every tissue (Supplementary Fig. 3a). For example, if a single nucleotide variant is synonymous on ENST1, a high-confidence LOFTEE stop-gained variant on ENST3 and ENST4, and low-confidence LOFTEE stop-gained variant on ENST5 and ENST6, the ext values will be synonymous: ENST1, stop-gained high-confidence: ENST3 + ENST4, and stop-gained low-confidence: ENST5 + ENST6 per tissue. This can be computed with the tx_annotate() function by setting the tx_annotation_type to 'expression'. We foresee the non-normalized ext values to be useful when only considering one tissue of interest.

To allow for taking average expression values across tissues of interest, we normalize the expression value for a given value to the total expression of the gene on which the variant is found. This is carried out by dividing the ext value with the sum of the expression of all transcripts per tissue in transcripts per million (TPM) (Supplementary Fig. 3b). The resulting pext value can be interpreted as the proportion of the total transcriptional output from a gene that would be affected by the given variant annotation in question. If the gene expression value (and thus the denominator) in a given tissue is 0, the pext value will not be available (NA) for that tissue.

When taking averages across tissues, such unavailable pext values are not considered (that is, when taking the mean across tissues, we remove NA values). This value can be computed with the tx_annotate() function by setting the tx_annotation_type to 'proportion'. For the analyses in this manuscript, we remove reproduction-associated GTEx tissues (endocervix, ectocervix, fallopian tube, prostate, uterus, ovary, testes and vagina), cell lines (transformed fibroblasts and transformed lymphocytes) and any tissue with less than 100 samples (bladder, brain Cervicalc-1 spinal cord, brain substantia nigra, kidney cortex and minor salivary gland), resulting in the use of 38 GTEx tissues.

We note that for a minority of genes, when RSEM[15] assigns higher relative expression to non-coding transcripts, the sum of the value of coding transcripts can be much smaller than the gene expression value for the transcript, resulting in low pext scores for all coding variants in the gene, and thus resulting in possible filtering of all variants for a given gene. In many cases this seems to be the result of spurious non-coding transcripts with a high degree of exon overlap with true coding transcripts. To prevent this artefact from affecting our analyses, we first calculated the maximum pext score for all variants across all protein-coding genes, and removed any gene where the maximum pext score was below 0.2. This resulted in the filtering of 668 genes, representing 3.3% of all genes analysed. We note that there is no overlap with the 668 genes and the haploinsufficient gene list, 97 of the filtered genes are present in OMIM (representing 1.5% of the OMIM gene list) and 42 filtered genes are considered constrained (representing 1.4% of LOEUF <0.35, or constrained, genes) thus having low effect on variant interpretation in the context of disease associations.

The full transcript-expression aware annotation pipeline, implemented in Hail 0.2, is fully available at https://github.com/macarthur-lab/tx_annotation with commands laid out for analyses in the manuscript. Passing a Hail table through the tx_annotate() function returns the same table with a new field entitled 'tx_annotation' which provides either the ext or pext value per variant-annotation pair, depending on parameter choice. We provide a helper function to extract the worst consequence and the associated expression values for these annotations. All analyses in the manuscript are based on the worst consequence of variant, ordered by VEP[38].

Functional validation of transcript-expression aware annotation

Conservation analysis was performed using phyloCSF scores using the same file used for the LOFTEE plugin, available publicly in gs://gnomad-public/papers/2019-tx-annotation/data/other_data/phylocsf_data.tsv.bgz . We denoted exons with a phyloCSF max open-reading frame score >1,000 as highly conserved and those with phyloCSF max open-reading frame score <−100 as lowly conserved (Supplementary Fig. 5a) and evaluated their average usage in GTEx.

Using the base-level pext values that are used in the gnomAD browser, we filtered to intervals with high or low conservation, and calculated the average pext value in the interval. To evaluate regions with low conservation but high expression, we identified genes harbouring unconserved regions with the pext value >0.9 for pathway enrichment analysis and used the web browser for FUMA GENE2FUNC feature[39], which incorporates Reactome[40], KEGG[41], Gene Ontology[42] (GO) as well as other ontologies. Default parameters were used for FUMA, with all protein coding genes as the background list. Results from FUMA pathway analysis are available in Supplementary Fig. 12, and full results are available in Supplementary Table 7.

Analysis of pext values for LOFTEE flags and the MAPS calculation were performed using the gnomAD v2.1.1 exome dataset. Calculation of MAPS scores was previously described[21] and is implemented as a Hail module, as also described previously[1]. MAPS is a relative metric, and cannot be compared across datasets, but is a useful summary metric for the frequency spectrum, indicating deleteriousness as inferred from rarity of variation (high values of MAPS correspond to lower frequency, suggesting the action of negative selection at more deleterious sites). The MAPS scores were calculated on the gnomAD v.2.1.1 dataset partitioning upon the LOEUF score and expression bin. The script for generating MAPs scores is available in the tx-annotation Github repository under /analyses/maps/maps_submit_per_class.py

Manual evaluation of unexpressed regions in haploinsufficient developmental delay genes using the GENCODE workflow

As an orthogonal evaluation of regions flagged as unexpressed with the pext metric, we identified any region in 61 haploinsufficient disease genes with a mean pext value <0.1 in all GTEx tissues and in GTEx brain samples, owing to the relevance of brain tissues for these disorders, regardless of mutational burden in gnomAD. The resulting list of 128 regions was evaluated by the HAVANA manual annotation group of the GENCODE project[20].

The manual evaluation first established whether the transcript model corresponding to the region in question was correct in terms of structure, comparing exon–intron combinations, and the accuracy of splice sites against the RNA evidence supporting the model. Second, the functional biotype of each model was reassessed; in particular, whether the decision to annotate the model as protein-coding in GENCODE v19 was appropriate. Note that GENCODE models that incorporate alternative exons or exon combinations in comparison to the 'canonical' isoform are likely to be annotated as coding if they contain a prospective CDS that is considered biologically plausible, based on a mechanistic view of translation. These re-annotations are summarized in Supplementary Table 5.

We binned cases into three main categories, according to confidence in both the accuracy and potential functional relevance of the overlapping models: (1) 'error', in which the model was seen to have an incorrect transcript structure and/or a CDS that conflicted with updated GENCODE annotation criteria (these annotations had been or will be changed in future GENCODE releases based on this evaluation); (2) 'putative', in which the model structure and CDS satisfied our current annotation criteria, although we judged the potential of the transcript represented to encode a protein with a functional role in cellular physiology to be nonetheless speculative (these have

been maintained as putative protein-coding transcripts in GENCODE); (3) 'validated', in which we believe it is highly probable that the model represents a true protein-coding isoform. High confidence in the validity of the CDS was based on comparative annotation, that is, the observation of CDS conservation and also the existence of equivalent transcript models in other species. GENCODE also annotates transcript models as 'nonsense-mediated decay' and 'non-stop decay', in which a translation is found that is predicted to direct the RNA molecule into cellular degradation programs. Although it has been established that such 'non-productive' transcription events can have a role in gene regulation and thus disease, the interpretation of variants within nonsense-mediated decay and non-stop decay CDS regions remains challenging[43]. These models were therefore classed in a separate category.

Gene list comparisons

To evaluate the filtering power of the pext metric for Mendelian variants, we evaluated the number of variants that would be filtered with an average GTEx pext cutoff of 0.1 (low expression) in the ClinVar and gnomAD datasets. We downloaded the ClinVar VCF from the ClinVar FTP (version dated 10/28/2018), imported it into Hail, annotated it with VEP v85 against Gencode v19, and added pext annotations with the tx_annotate() function. All evaluated variants were annotated as HC by LOFTEE v1.0, and ClinVar variants were filtered to those marked as pathogenic, with no conflicts, and reviewed with at least one star status.

For variants in 61 haploinsufficient genes, we identified any variant identified in at least one individual with any zygosity in both datasets. For variants identified in autosomal recessive disease genes, we used a list of 1,183 OMIM disease genes deemed to follow a recessive inheritance pattern by Blekhman et al.[44] and Berg et al.[45] (available as https://github.com/macarthur-lab/gene_lists/blob/master/lists/all_ar.tsv). We compared the pext value for all pLoF variants identified in ClinVar versus any variant in a homozygous state in at least one individual in the gnomAD exome or genome datasets. Finally, we used a LOEUF cutoff of 0.35 to denote constrained genes, and compared any synonymous or pLoF variant in these genes in the gnomAD exome or genome datasets.

De novo and rare variant analysis

De novo variants were collated from previously published studies. We collected de novo variants identified in 5,305 probands from trio studies of intellectual disability/developmental disorders (Hamdam et al.[27]: $n = 41$, de Ligt et al.[28]: $N = 100$, Rauch et al.[29]: $N = 51$, DDD[24]: $n = 4{,}293$, Lelieveld et al.[26]: $n = 820$), 1,073 probands with congenital heart disease with co-morbid developmental delay (Sifrim et al.[46]: $n = 512$, Chih Jin et al.[47]: 561), 6,430 ASD probands, and 2,179 unaffected controls from the Autism Sequencing Consortium[25]. We also used a previously published dataset of variants in 8,437 cases with ASD and/or attention-deficit/hyperactivity disorder and 5,214 controls from the Danish Neonatal Screening Biobank[48]. In this analysis, we analysed pLoF variants identified in highly constrained genes (first LOEUF decile) with a combined total allele count of ≤ 10 in cases and controls.

We annotated both de novo and rare variants with VEP v85 against Gencode v19 and added pext annotations with the tx_annotate() function. We then calculated the average pext metric across 11 GTEx brain samples and binned them as low (pext < 0.1), medium (0.1 ≤ pext ≤ 0.9) or high (pext > 0.9) expression. We then calculated the number of pLoF, missense, and synonymous variants per pext expression bin. To obtain case-control rate ratios and the 95% confidence intervals for de novo variant analyses, we used a two-sided Poisson exact test on counts. To obtain the odds ratio for the rare variant analysis in ASD/ADHD, we used the Fisher's exact test for count data.

Isoform quantifications via salmon

To evaluate whether use of a different isoform quantification tool would affect results, we compared results of *TCF4* base-level expression (shown in Fig. 2b), MAPS (Fig. 3c) and comparison of the number of variants filtered in haploinsufficient developmental disease genes in ClinVar vs gnomAD (Fig. 4a) using RSEM quantifications used in this study with quantifications using salmon v.0.12[17]. Due to the intractability of re-quantifying the entire GTEx dataset, we downloaded and requantified 151 GTEx brain cortex CRAM files from the V7 dataset. We first converted CRAMs to fastq files using Picard 2.18.20 and ran salmon with the 'salmon quant –i index -fastq1 – fastq2 –minAssignedFrag1 –validateMappings' command. The index was created with the 'salmon index –t transcript.fa –type quasi –k 31' command using the GENCODE v19 protein-coding and lncRNA transcripts FASTA files. The existing GTEx RSEM isoform quantifications were filtered to the same GTEx brain cortex samples. For the analyses to remain consistent with the remainder of the manuscript, we calculated the maximum brain cortex pext score for all variants across all protein-coding genes for both the RSEM and salmon quantifications, and removed any gene in which the maximum pext score was below 0.2. This resulted in filtering 325 genes from the salmon quantification of the brain cortex samples and 691 genes from the RSEM quantification, corresponding to 3.4 and 1.6% of quantified genes, respectively. We filtered these genes in both the MAPs and gene list comparison analysis seen in Supplementary Fig. 11. The WDL script for the quantification pipeline is available at: gs://gnomad-public/papers/2019-tx-annotation/results/salmon_rsem/salmon.wdl and the commands to obtain results for each individual analysis in the tx-annotation Github repository under /analyses/rsem_salmon/.

Transcript expression aware annotation with fetal isoform expression dataset

Although our analyses were based on transcript expression aware annotation from the GTEx v7 dataset, we provide necessary files for pext annotation with the Human Brain Development Resource (HBDR) fetal brain dataset[49] in gs://gnomad-public/papers/2019-tx-annotation/data/HBDR_fetal_RNaseq. HBDR includes 558 samples from varying brain subregions across developmental time points. We downloaded HDBR sample fastq files from European Nucleotide Archive (study accession PRJEB14594) and obtained RSEM isoform quantification on HBDR fastqs using the GTEx v7 quantification pipeline, publicly available at https://github.com/broadinstitute/gtex-pipeline/) which briefly involves two-pass alignment with STAR v2.4.2a[50] and isoform quantification with RSEM v1.2.22. Here, we also removed genes where the average pext across HBDR was below 0.2, resulting in the removal of 712 genes (3.5% of all analysed genes). The dataset was also used for the analysis of baselevel expression values in *SCN2A* shown in Supplementary Fig. 7d.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

We used the gnomAD v.2.1.1 sites Hail 0.2 (https://hail.is) table that is accessible publicly at gs://gnomad-public/release/2.1.1 and at https://gnomad.broadinstitute.org. The GTEx v7 gene and isoform expression data were downloaded from the GTEx portal (gtexportal.org). The LOEUF constraint file was downloaded from gs://gnomad-resources/lof_paper/. All files used in the analyses in the manuscript are available in gs://gnomad-public/papers/2019-tx-annotation/.

Code availability

The GTEx pipeline for isoform quantification is available publicly (https://github.com/broadinstitute/gtex-pipeline/) and briefly involves two-pass alignment with STAR v2.4.2a[50], gene expression quantification

with RNA-SeQC v1.1.8[51], and isoform quantification with RSEM v1.2.22. Variants used in all gnomAD analyses in the manuscript passed random forest filtering, and all pLoF variants were annotated as high-confidence by LOFTEE v.1.0, which is described in an accompanying manuscript[1]. Scripts to quality control the gnomAD dataset are available at https://github.com/macarthur-lab/gnomad_qc and the scripts to generate files for the analyses are available at https://github.com/macarthur-lab/tx_annotation.

34. Riggs, E. R. et al. Copy number variant discrepancy resolution using the ClinGen dosage sensitivity map results in updated clinical interpretations in ClinVar. *Hum. Mutat.* **39**, 1650–1659 (2018).
35. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
36. Kent, W. J. et al. The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
37. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
38. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
39. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
40. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
41. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
42. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
43. da Costa, P. J., Menezes, J. & Romão, L. The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. *Int. J. Biochem. Cell Biol.* **91** (Pt B), 168–175 (2017).
44. Blekhman, R. et al. Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18**, 883–889 (2008).
45. Berg, J. S. et al. An informatics approach to analyzing the incidentalome. *Genet. Med.* **15**, 36–44 (2013).
46. Sifrim, A. et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–1065 (2016).
47. Jin, S. C. et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
48. Satterstrom, F. K. et al. ASD and ADHD have a similar burden of rare protein-truncating variants. Preprint at https://www.bioRxiv.org/content/10.1101/277707v1 (2018).
49. Lindsay, S. J. et al. HDBR expression: a unique resource for global and individual gene expression studies during early human brain development. *Brain Dev.* **10**, 86 (2016).
50. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
51. DeLuca, D. S. et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).

Acknowledgements We thank all of the research participants for contributing their data. This work was supported by NIDDK U54 DK105566, NIGMS R01 GM104371, and the Broad Institute. KJK was supported by NIGMS F32 GM115208. A.O.L was supported by NICHD K12 HD052896. The GENCODE project is supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U41HG007234. The results published here are in part based on data: (1) generated by The Cancer Genome Atlas (TCGA) managed by the NCI and NHGRI (accession: phs000178.v10.p8); information about TCGA can be found at http://cancergenome.nih.gov; (2) generated by the Genotype-Tissue Expression Project (GTEx) managed by the NIH Common Fund and NHGRI (accession: phs000424.v7.p2); (3) generated by the Exome Sequencing Project, managed by NHLBI; and (4) generated by the Alzheimer's Disease Sequencing Project (ADSP), managed by the NIA and NHGRI (accession: phs000572.v7.p4). We thank E. Pierce-Hoffman for previous analysis and thoughts on characterizing loss-of-function variants in haploinsufficient genes. We thank the iPSYCH/SSI/Broad Institute psychiatric genetics study for the use of exome count data. We have complied with all relevant ethical regulations. This study was overseen by the Broad Institute's Office of Research Subject Protection and the Partners Human Research Committee, and was given a determination of Not Human Subjects Research. Informed consent was obtained from all participants. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions B.B.C., K.J.K., J.A. and D.G.M. contributed to the writing of the manuscript and generation of figures. B.B.C., K.J.K., E.G.S., M.S.-B., J.M.M., J.A.K., M.J.D. and D.G.M. contributed to the analysis of data. B.B.C., K.J.K., J.A.K., J.M.M., J.K., F.K.S., A.H.O'D.-L., T.P., C.S., M.J.D. and D.G.M. developed tools and methods that enabled the scientific discoveries herein. B.B.C., K.J.K., N.A.W., T.P., M.S., J.A., M.J.D. and D.G.M. contributed to the production and quality control of the gnomAD dataset. All authors listed under The Genome Aggregation Database Consortium contributed to the generation of the primary data incorporated into the gnomAD resource. All authors reviewed the manuscript.

Competing interests K.J.K. owns stock in Personalis. A.H.O'D.-L. has received honoraria from ARUP and Chan Zuckerberg Initiative. M.J.D. is a founder of Maze Therapeutics. D.G.M. is a founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2329-2.

Correspondence and requests for materials should be addressed to D.G.M.

Peer review information *Nature* thanks Roderic Guigo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.

nature research

Corresponding author(s): Beryl B Cummings
Daniel G MacArthur

Last updated by author(s): Feb 20, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for the collection of data, as this was an opportunistic study.
Data analysis	All code to perform quality control and data analysis is provided in the following Github repos: https://github.com/macarthur-lab/tx_annotation https://github.com/macarthur-lab/gnomad_qc https://github.com/macarthur-lab/gnomad_lof https://github.com/konradjk/loftee https://github.com/broadinstitute/gtex-pipeline/ Hail 0.2 is available at: https://hail.is/ RSEM v1.2.22 : https://deweylab.github.io/RSEM/ RNA-SeQC v1.1.8 : https://github.com/broadinstitute/rnaseqc STAR v2.4.2a : https://github.com/alexdobin/STAR Alamut v.2.11 https://www.interactive-biosoftware.com/alamut-visual/ Variant Effect Predictor (VEP) v.85 : https://uswest.ensembl.org/info/docs/tools/vep/index.html FUMA GENE2FUNC v1.3.5e : https://fuma.ctglab.nl/ R version 3.4.0 : https://cran.r-project.org/bin/macosx/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets are described in the manuscript or Supplementary Information, including deposition of the full dataset at https://gnomad.broadinstitute.org. Data for specific analyses are available publicly at gs://gnomad-public/papers/2019-tx-annotation/ and the specific folders therein for analyses are referenced in the manuscript for ease of recreating analyses with the data provided. There are no restrictions on the aggregate data released.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study provides a framework and tool to improve variant interpretation in datasets of any size, no matter how small or large. As a proof of principle, we use one of the largest datasets of human genetic variation, gnomAD and the largest functional genomics dataset GTEx. In other words, this study is opportunistic, and involves secondary use of available genome, exome and transcriptome data. No sample size was predetermined.
Data exclusions	Sample QC and variant QC for gnomAD are described extensively in the supplementary methods of the main manuscript. Notably, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded. For the analyses in the manuscript, we removed GTEx tissues with low sample numbers, reproductive tissue and non-tissues (ie. cell lines). For the purpose of our manuscript, we did not define pre-exclusion criteria for calculation of pext. However for the analyses in the manuscript, we defined pre-exclusion tissues : we removed reproduction-associated GTEx tissues (endocervix, ectocervix, fallopian tube, prostate, uterus, ovary, testes, vagina), cell lines (transformed fibroblasts, transformed lymphocytes) and any tissue with less than one hundred samples (bladder, brain Cervicalc-1 spinal cord, brain substantia nigra, kidney cortex, minor salivary gland) . This is explained in the Methods section of the manuscript.
Replication	We did not attempt to reproduce any findings in a separate but identical dataset, as no other data set of comparable size exists. However we replicate key findings with a seperate isoform quantification tool. We successfully replicate the MAPS results, shown in Supplementary Figure 11. We also use an external fetal dataset to provide additional data.
Randomization	As this was a population-based study, and not a case-control study, no randomization was performed.
Blinding	As this was a population-based study, and not a case-control study, blinding was not relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	As an opportunistic collection of data, the participants in this study were not selected based on age, gender, or genotypic information. As described above, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded from gnomAD. As an opportunistic collection of data, the participants in gnomAD were not selected based on age, gender, or genotypic information. The populations are provided in Supplementary Table 7 of the accompanying Karczewski et al., and there are 64,754 females and 76,702 males. These data were obtained primarily from case-control studies of adult-onset common diseases, including cardiovascular disease, type 2 diabetes, and psychiatric disorders. GTEx v7 collection was similarly opportunistic and has been previously extensively published and reported on. Population characteristics of the data can be found in Reference 2 : GTEx Consortium et al., Genetic effects on gene expression across human tissues. Nature 550, 204 (2017).
Recruitment	As this was an opportunistic secondary use study, we did not recruit any participants.
Ethics oversight	This study was overseen by the Broad Institute's Office of Research Subject Protection and the Partners Human Research Committee, and was given a determination of Not Human Subjects Research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Evaluating drug targets through human loss-of-function genetic variation

https://doi.org/10.1038/s41586-020-2267-z

Received: 25 January 2019

Accepted: 10 February 2020

Published online: 27 May 2020

Open access

Check for updates

Eric Vallabh Minikel[1,2,3,4,5,6,7,8]✉, Konrad J. Karczewski[1,4], Hilary C. Martin[9], Beryl B. Cummings[1,4,5], Nicola Whiffin[1,10], Daniel Rhodes[11], Jessica Alföldi[1,4], Richard C. Trembath[12], David A. van Heel[13], Mark J. Daly[1,4], Genome Aggregation Database Production Team*, Genome Aggregation Database Consortium*, Stuart L. Schreiber[3,14] & Daniel G. MacArthur[1,4,153,154]✉

Naturally occurring human genetic variants that are predicted to inactivate protein-coding genes provide an in vivo model of human gene inactivation that complements knockout studies in cells and model organisms. Here we report three key findings regarding the assessment of candidate drug targets using human loss-of-function variants. First, even essential genes, in which loss-of-function variants are not tolerated, can be highly successful as targets of inhibitory drugs. Second, in most genes, loss-of-function variants are sufficiently rare that genotype-based ascertainment of homozygous or compound heterozygous 'knockout' humans will await sample sizes that are approximately 1,000 times those presently available, unless recruitment focuses on consanguineous individuals. Third, automated variant annotation and filtering are powerful, but manual curation remains crucial for removing artefacts, and is a prerequisite for recall-by-genotype efforts. Our results provide a roadmap for human knockout studies and should guide the interpretation of loss-of-function variants in drug development.

Human genetics is an increasingly crucial source of evidence guiding the selection of new targets for drug discovery[1]. Most new clinical drug candidates eventually fail for lack of efficacy[2], and although in vitro, cell culture and animal model systems can provide preclinical evidence that the compound engages its target, too often the target itself is not causally related to human disease[1]. Candidates targeting genes with human genetic evidence for disease causality are more likely to reach approval[3,4], and identification of humans with loss-of-function (LoF) variants, particularly two-hit (homozygous or compound heterozygous) genotypes, has, for several genes, correctly predicted the safety and phenotypic effect of pharmacological inhibition[5]. Although these examples demonstrate the value of human genetics in drug development, important questions remain regarding strategies for identifying individuals with LoF variants in a gene of interest, interpretation of the frequency—or lack—of such individuals, and whether it is wise to pharmacologically target a gene in which LoF variants are associated with a deleterious phenotype.

Public databases of human genetic variation have catalogued predicted loss-of-function (pLoF) variants—nonsense, essential splice site, and frameshift variants expected to result in a non-functional allele. This presents an opportunity to study the effects of pLoF variation in genes of interest and to identify individuals with pLoF genotypes to understand gene function or disease biology, or to assess potential for therapeutic targeting. Although many variants initially annotated as pLoF do not, in fact, abolish gene function[6], rigorous automated filtering can remove common error modes[7]. True LoF variants are generally rare, and show important differences between outbred, bottlenecked[8] and consanguineous[9] populations[6,10]. Counting the number of distinct pLoF variants in each gene in a population sample allows the quantification of gene essentiality in humans through a metric named 'constraint'[10–13]. Specifically, the rate at which de novo pLoF mutations arise in each gene is predicted on the basis of rates of DNA mutation[10,12], and the ratio of the count of pLoF variants observed in a database to the number expected based on mutation rates—obs/exp, or constraint score—measures how strongly purifying natural selection has removed such variants from the population. The annotation of pLoF variants remains imperfect, and continued improvements are being made[14], but constraint usefully measures gene essentiality, as demonstrated by agreement with cell culture and mouse knockout experiments[7], by overlap with human disease genes[7,10] and genes depleted for structural variation[15], and by the power of constraint to enrich for deleterious variants in neurodevelopmental disorders[7,16].

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Chemical Biology and Therapeutics Science Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [5]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. [6]Henry and Allison McCance Center for Brain Health, Massachusetts General Hospital, Boston, MA, USA. [7]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. [8]Prion Alliance, Cambridge, MA, USA. [9]Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK. [10]National Heart and Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. [11]Centre for Translational Bioinformatics, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London and Barts Health NHS Trust, London, UK. [12]School of Basic and Medical Biosciences, Faculty of Life Sciences and Medicine, King's College London, London, UK. [13]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [14]Department of Chemistry & Chemical Biology, Harvard University, Cambridge, MA, USA. [153]Present address: Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, Australia. [154]Present address: Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: eminikel@broadinstitute.org; d.macarthur@garvan.org.au

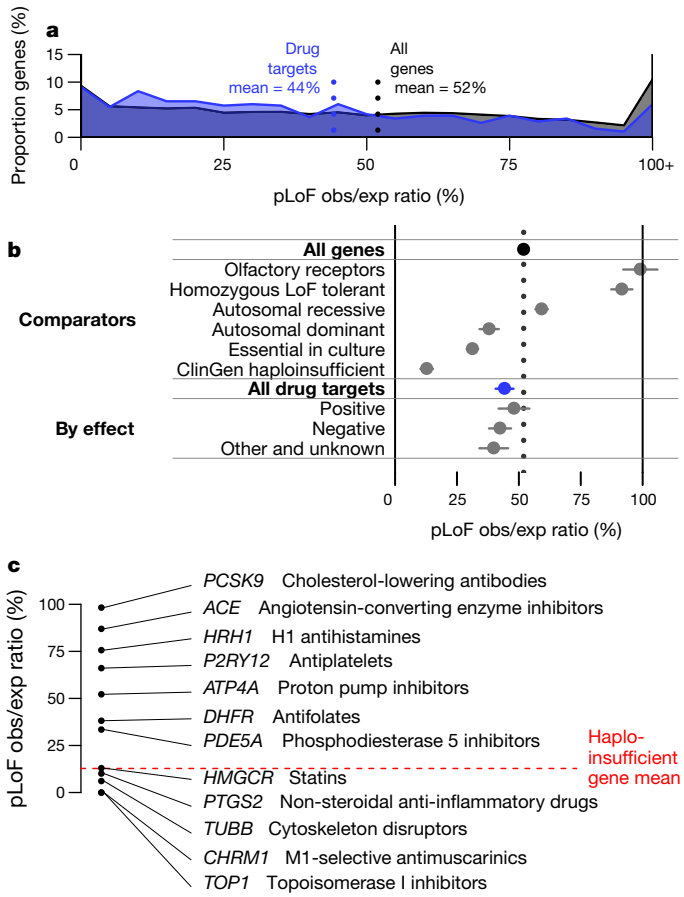


Fig. 1 | pLoF constraint in drug targets. a, Histogram of pLoF obs/exp values for all genes (black, $n = 17{,}604$) versus drug targets (blue, $n = 383$). **b**, Forest plot of means (dots) and 95% confidence intervals of the mean (line segments), for constraint in the indicated gene sets (data sources and n values in Extended Data Table 1). For drug effect, 'positive' indicates agonist, activator or inducer, whereas negative indicates antagonist, inhibitor or suppressor, for example. **c**, Examples of drug targets and corresponding drug classes from across the constraint spectrum. Details in Extended Data Table 2.

Building on these insights, here we leverage pLoF variation in the Genome Aggregation Database (gnomAD)[7] v2 dataset of 141,456 individuals to answer open questions in the interpretation of human pLoF variation in disease biology and drug development.

Constraint in human drug targets

We compared constraint in the targets of approved drugs extracted from DrugBank[17] ($n = 383$) versus all protein-coding genes ($n = 17{,}604$). Drug targets were, on average, just slightly more constrained than all genes (mean 44% versus 52%, nominal $P = 0.00028$, $D = 0.11$, two-sided Kolmogorov–Smirnov test), but the two gene sets had a qualitatively similar distribution of scores, ranging from intensely constrained (0% obs/exp) to not at all constrained (≥100% obs/exp) (Fig. 1a). Constraint scores showed clear divergence between categories of genes (Extended Data Table 1) expected to be more or less tolerant of inactivation (Fig. 1b), as previously reported[7,10], validating the usefulness of constraint as a measure of gene essentiality. Nonetheless, when drug targets were stratified by drug effect (Fig. 1b), modality, or indication (Extended Data Fig. 1), no statistically significant differences between subsets of drug targets were observed.

The slightly but significantly lower obs/exp value among drug targets may superficially appear to provide evidence that constrained genes make superior drug targets. Stratification of drug targets by protein family, human disease association, and tissue expression, however, argues against this interpretation. Drug targets are strongly enriched for a few canonically 'druggable' protein families, for genes known to be involved in human disease, and for genes with tissue-restricted expression; each of these properties is in turn correlated with either significantly stronger or weaker constraint (Extended Data Fig. 2). Although controlling for these correlations does not abolish the trend of stronger constraint among drug targets, the correlation of so many observed variables with the status of a gene as a drug target argues that many unobserved variables probably also confound interpretation of the lower mean obs/exp value among drug targets.

The overall constraint distribution of drug targets (Fig. 1a) also argues against the view that a gene in which LoF is associated with a deleterious phenotype cannot be successfully targeted. Indeed, 19% of drug targets ($n = 73$), including 52 targets of inhibitors, antagonists or other 'negative' drugs, have lower obs/exp values than the average (12.8%) for genes known to cause severe diseases of haploinsufficiency[18] (ClinGen level 3). To determine whether this finding could be explained by a particular class or subset of drugs, we examined constraint in several well-known example drug targets (Fig. 1c, Extended Data Table 2). Some heavily constrained genes are targets of cytotoxic chemotherapy agents such as topoisomerase inhibitors or cytoskeleton disruptors, a set of drugs intuitively expected to target essential genes. However, genes with near-complete selection against pLoF variants also include *HMGCR* and *PTGS2*, the targets of highly successful, chronically used inhibitors—statins and aspirin.

These human in vivo data further the evidence from other species and models that essential genes can be good drug targets. Homozygous knockout of *Hmgcr* and *Ptgs2* are lethal in mice[19–21]. Drug targets exhibit higher inter-species conservation than other genes[22]. Targets of negative drugs include 14 genes with lethal heterozygous knockout mouse phenotypes reported[23] and 6 reported as essential in human cell culture[24].

Prospects for finding human 'knockouts'

Athough constraint alone is not adequate to nominate or exclude drug targets, the study of individuals with single hit (heterozygous) or two-hit ('knockout') LoF genotypes in a gene of interest can be highly informative about the biological effect of engaging that target[5]. To assess prospects for ascertaining knockout individuals, we computed the cumulative allele frequency (CAF) of pLoF variants in each gene (Methods), and then used this to estimate the expected frequency of two-hit individuals under different population structures (Fig. 2) in the absence of natural selection.

Whereas gnomAD is now large enough to include at least one pLoF heterozygote for most (15,317 out of 19,194; 79.8%) genes, ascertainment of total knockout individuals in outbred populations will require 1,000-fold larger sample sizes for most genes: the median expected two-hit frequency of a gene is just six per billion (Fig. 2a). Even if every human on Earth were sequenced, there are 4,728 genes (24.6%) for which identification of even one two-hit individual would not be expected in outbred populations. Intuitively, because the sample size of gnomAD today is larger than the square root of the world population, variants so far seen in zero or only a few heterozygous individuals are not likely to ever be seen in a homozygous state in outbred populations, except where variants prove common in populations not yet well-sampled by gnomAD.

Because population bottlenecks can result in very rare variants present in a founder rising to an unusually high frequency, we also considered knockout discovery in bottlenecked populations, using Finnish individuals in gnomAD as an example[8]. Although this population structure can enable well-powered association studies for the small fraction of genes in which pLoF variants drifted to high frequency due to the bottleneck, overall, identification of two-hit pLoF individuals

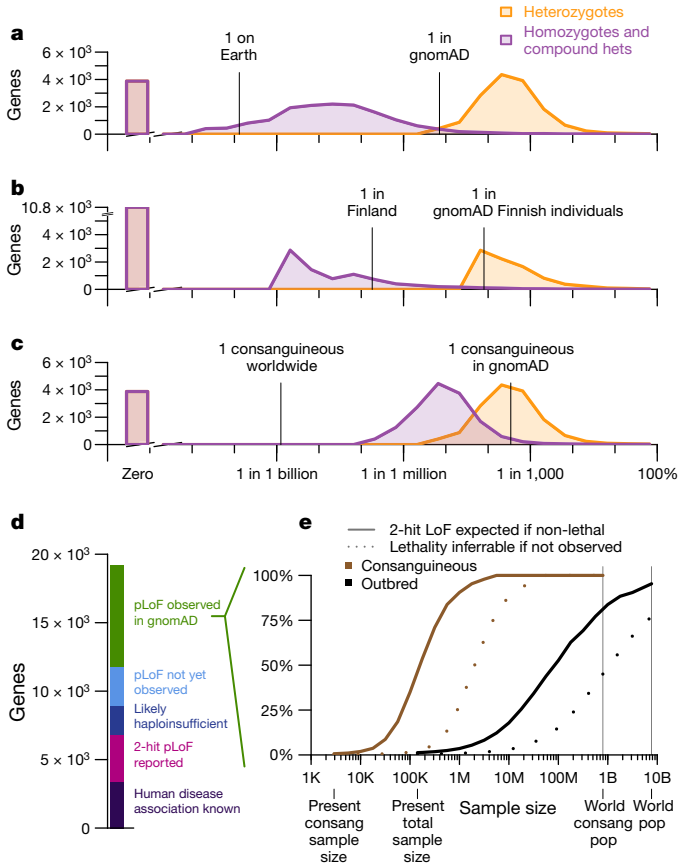


Fig. 2 | Prospects for discovery of human knockouts. a–**c**, Histograms (**a**–**c**): genes by expected heterozygote frequency (orange), and two-hit homozygote and compound heterozygote frequency (purple). **a**, Outbred populations. **b**, Finnish individuals; an example of a bottlenecked population. **c**, Consanguineous individuals. **d**, Current status of pLoF or disease association discovery for all protein-coding genes. **e**, Projected sample sizes required for discovery of two-hit individuals (solid lines) and for statistical inference that a two-hit genotype is lethal if no such individuals are observed (dashed lines), for 'pLoF observed in gnomAD' genes (**d**) for consanguineous and outbred individuals.

for a pre-specified gene of interest appears equally or more difficult in Finnish individuals than in outbred populations (Fig. 2b, Extended Data Fig. 3), because rare variants not present in a founder have been effectively removed from the population.

In consanguineous individuals, parental relatedness greatly increases the frequency of homozygous pLoF genotypes. The $n = 2{,}912$ individuals in the East London Genes & Health (ELGH) cohort[25] who report having parents who are second cousins or closer have on average 5.8% of their genomes autozygous. Here, the expected frequency of two-hit individuals is many times higher than in outbred populations, at five per million for the median gene (Fig. 2c).

These projections allow us to draft a roadmap for discovery of human knockouts across 19,194 genes (Fig. 2d, e). Online Mendelian Inheritance in Man (OMIM) already describes human disease association for 3,367 genes (18%), although the discovery of LoF individuals in population databases will still be valuable for assessing penetrance and identifying LoF syndromes of known gain-of-function genes. Another 3,421 genes (18%) without known human disease association have two-hit pLoF genotypes reported in gnomAD[7], ELGH[26], PROMIS[27], deCODE[28] or UK Biobank[29], which suggests that this genotype may be tolerated. An additional 2,190 genes (11%) appear intolerant of heterozygous inactivation (pLI score > 0.9) in gnomAD—a set expected to be enriched for genes with severe heterozygous and lethal homozygous LoF phenotypes. Another 2,781 genes (14%) have no pLoF variants observed in gnomAD, but our sample size is not yet large enough to robustly infer LoF intolerance. For these genes, observation of outbred two-hit individuals is not expected, and we cannot yet assess the feasibility of identifying consanguineous two-hit individuals because we lack an estimate of pLoF allele frequency.

This leaves 7,435 genes (39%) for which one or more pLoFs are observed in gnomAD, but strong LoF intolerance cannot be determined, two-hit genotypes have not been observed, and a human disease phenotype is not known. We projected the sample sizes required to identify knockout individuals for these genes (Fig. 2e). In outbred populations, current sample sizes would need to increase by approximately 1,000-fold before ascertainment of a single two-hit LoF individual would be expected for the typical gene. By contrast, around a 10- to 100-fold increase from current consanguineous sample size, meaning hundreds of thousands of individuals in absolute terms, would identify at least one two-hit LoF individual for the typical gene. Among other simplifying assumptions (Methods), these projections presume that complete knockout is tolerated. When only one or a few two-hit individuals are expected in a dataset, the absence of any such individuals can be due to either early lethality, a severe clinical phenotype incompatible with inclusion in gnomAD, or simply chance. Thus, the ability to infer lethality of the two-hit genotype based on statistical evidence will lag behind the identification of two-hit individuals where they do exist (Fig. 2e). For some genes, inference of lethality will always remain impossible in outbred populations, though it may be feasible in consanguineous individuals.

Curation of pLoF variants

Where pLoF variants can be identified, they are a valuable resource for assessing the effect of lifelong reduction in gene dosage. To highlight the challenges and opportunities of identifying such variants, we manually curated gnomAD data and the scientific literature for six genes associated with gain-of-function (GoF) neurodegenerative diseases, for which inhibitors or suppressors are under development[30–35]: *HTT* (Huntington's disease), *MAPT* (tauopathies), *PRNP* (prion disease), *SOD1* (amyotrophic lateral sclerosis), and *LRRK2* and *SNCA* (Parkinson's disease). The results (Fig. 3, Extended Data Table 3) illustrate four points about pLoF variant curation.

First, other things being equal, genes with longer coding sequences offer more opportunities for LoF variants to arise, and so tend to have a higher cumulative frequencies of LoF variants, unless they are heavily constrained. Ascertainment of LoF individuals is thus harder for shorter and/or more constrained genes, even though these may be good targets.

Second, many variants annotated as pLoF are false positives[6], and these are enriched for higher allele frequencies, so that both filtering and curation have an outsized effect on the cumulative allele frequency of LoF. Studies of human pLoF variants lacking stringent curation can therefore easily dilute results with false pLoF carriers.

Third, after careful curation, cumulative LoF allele frequency is sometimes sufficiently high to place certain bounds on what heterozygote phenotype might exist. For example, GoF mutations causing genetic prion disease have a genetic prevalence of approximately 1 in 50,000[36] and have been known for three decades, with thousands of cases identified, making it unlikely that a comparably severe and penetrant haploinsufficiency syndrome associated with *PRNP* would have gone unnoticed to the present day despite being more than twice as common (roughly 1 in 18,000). Similar arguments can be made for *HTT*, *LRRK2* and *SOD1* genes (Extended Data Tables 3, 4). Of course, this does not rule out a less severe or less penetrant heterozygous LoF phenotype.

Finally, careful inspection of the distributions of pLoF variants can reveal important error modes or disease biology. *HTT*, *MAPT*

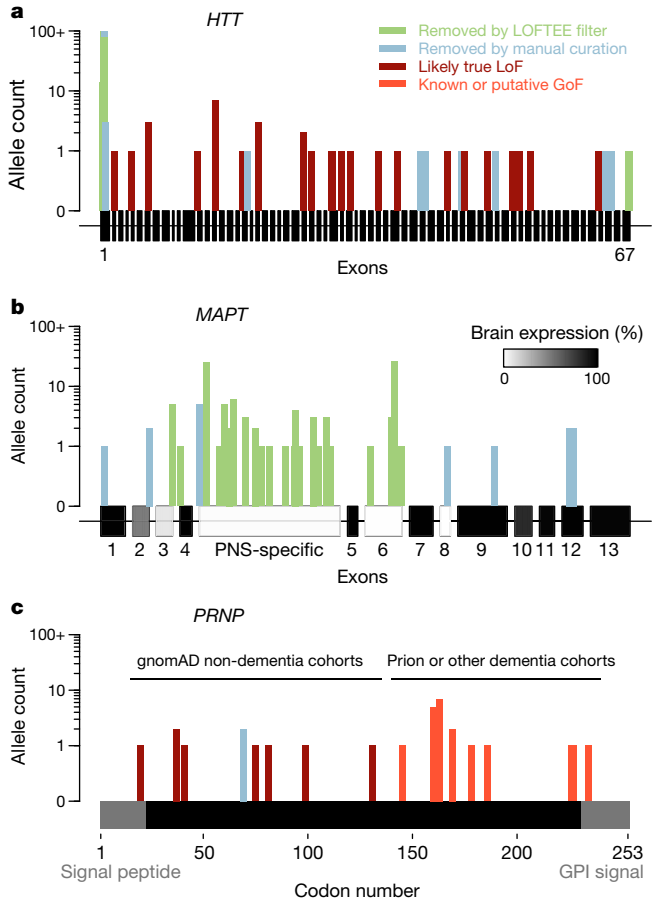


Fig. 3 | Insights from non-random positional distributions of pLoF variants. **a**–**c**, *HTT* (**a**), *MAPT*, with brain expression data from GTEx[40] (**b**) and *PRNP*, a single protein-coding exon with domains removed by post-translational modification in grey (**c**), showing previously reported variants[41] and those newly identified in gnomAD and in the literature (Extended Data Table 5). GPI, glycosylphosphatidylinositol. Detailed variant curation results are provided in Supplementary Table 1.

and *PRNP* genes each have different non-random positional distributions of pLoF variants (Fig. 3). High-frequency *HTT* pLoF variants cluster in the polyglutamine/polyproline repeat region of exon 1 and appear to be alignment artefacts (Fig. 3a). True *HTT* LoF variants are rare and the gene is highly constrained, which might suggest some fitness effect in a heterozygous state in addition to the known severe homozygous phenotype[37,38], although the frequency of LoF carriers still argues against a penetrant syndromic illness, consistent with the lack of phenotype reported in heterozygotes identified so far[38,39]. High-frequency *MAPT* pLoF variants cluster in exons not expressed in the brain in GTEx data[14,40], and all remaining pLoFs appear to be alignment or annotation errors (Fig. 3b). No true LoFs are observed in *MAPT*, although our sample size is insufficient to prove that *MAPT* LoF is not tolerated—among constitutive brain-expressed exons, we expect 12.6 LoFs and observe 0, giving a 95% confidence interval upper bound of 23.7% for obs/exp values. *PRNP*-truncating variants in gnomAD cluster in the N terminus; the sole C-terminal truncating variant in gnomAD is a dementia case (Extended Data Table 5), consistent with variants at codon ≥145 causing a pathogenic gain-of-function through change in localization (Fig. 3c). Within codons 1–144, *PRNP* is unconstrained (Extended Data Table 3), and no neurological phenotype has been identified in individuals with truncating variants so far, consistent with the hypothesis that N-terminal truncating variants are true LoF and are tolerated in a heterozygous state[41].

Discussion

Studying human gene inactivation can illuminate human biology and guide the selection of drug targets, complementing mouse knockout studies[42], but analysis of any one gene requires genome-wide context to set expectations and guide inferences. Here we have used gnomAD data to provide context to aid in the interpretation of human LoF variants.

Targets of approved drugs range from highly constrained to completely unconstrained. There may be several reasons why some genes apparently tolerate pharmacological inhibition but not genetic inactivation. LoF variants in constitutive exons should affect all tissues for life, whereas drugs differ in tissue distribution and timing and duration of use. Many drugs known or suspected to cause fetal harm are tolerated in adults[43], and might target developmentally important genes. Constraint is thought to primarily reflect selection against heterozygotes[13], the effective gene dosage of which may differ from that achieved by a drug. Constraint measures natural selection over centuries or millennia; the environment of our ancestors presented different selective pressures from what we face today. Actions of small-molecule drugs may not map one-to-one onto genes[44–47]. Regardless, these human in vivo data show that even a highly deleterious knockout phenotype is compatible with a gene being a viable drug target.

For most genes, the lack of total knockout individuals identified so far does not yet provide statistical evidence that this genotype is not tolerated. Indeed, for many genes, such evidence may never be attainable in outbred populations. Bottlenecked populations, individually, are unlikely to yield two-hit individuals for a pre-specified gene of interest, although the sequencing of many different, diverse bottlenecked populations will certainly expand the set of genes accessible by this approach. Identification of two-hit individuals will be most greatly aided by increased investment in consanguineous cohorts, in which the sample size required for any given gene is often orders of magnitude lower than in outbred populations. Our analysis is limited by sample size, insufficient diversity of sampled populations, and simplifying assumptions about population structure and distribution of LoF variants, so our calculations should be taken as rough, order-of-magnitude estimates. Nonetheless, this strategic roadmap for the identification of human knockouts should inform future research investments and rationalize the interpretation of existing data.

Recall-by-genotype efforts are only valuable if the variants in question are correctly annotated. Automated filtering[7] and transcript expression-aware annotation[14] are powerful tools, but we demonstrate the continued value of manual curation for excluding further false positives, assessing and interpreting the cumulative allele frequency of true LoF variants, and identifying error modes or biological phenomena that give rise to non-random distributions of pLoF variants across a gene. Such curation is essential before any recontact efforts, and establishing methods for high-throughput functional validation[48] of LoF variants is a priority. Our curation of pLoF variants in neurodegenerative disease genes is limited by a lack of functional validation and detailed phenotyping; a companion paper demonstrates a deeper investigation of the effects of LoF variants in the *LRRK2* gene[49].

Drug development projects may increasingly be accompanied by efforts to phenotype human carriers of LoF variants. With the cost of drug discovery driven overwhelmingly by failure[50], successful interpretation of LoF data to select the right targets and right clinical pathways will yield outsize benefits for research productivity and, ultimately, human health.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2267-z.

1. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
2. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32**, 40–51 (2014).
3. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
4. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
5. Musunuru, K. & Kathiresan, S. Genetics of common, complex coronary artery disease. *Cell* **177**, 132–145 (2019).
6. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
7. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* https://doi.org/10.1038/s41586-020-2308-7 (2020).
8. Lim, E. T. et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10**, e1004494 (2014).
9. Bittles, A. H. & Black, M. L. Evolution in health and medicine Sackler colloquium: consanguinity, human evolution, and complex diseases. *Proc. Natl Acad. Sci. USA* **107** (Suppl. 1), 1779–1786 (2010).
10. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
11. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
12. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
13. Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
14. Cummings, B. B. et al. Transcript expression-aware annotation improves rare variant discovery and interpretation. *Nature* https://doi.org/10.1038/s41586-020-2329-2 (2020).
15. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* https://doi.org/10.1038/s41586-020-2287-8 (2020).
16. Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017).
17. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46** (D1), D1074–D1082 (2018).
18. Rehm, H. L. et al. ClinGen—the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
19. Morham, S. G. et al. Prostaglandin synthase 2 gene disruption causes severe renal pathology in the mouse. *Cell* **83**, 473–482 (1995).
20. Ohashi, K. et al. Early embryonic lethality caused by targeted disruption of the 3-hydroxy-3-methylglutaryl-CoA reductase gene. *J. Biol. Chem.* **278**, 42936–42941 (2003).
21. Nagashima, S. et al. Liver-specific deletion of 3-hydroxy-3-methylglutaryl coenzyme A reductase causes hepatic steatosis and death. *Arterioscler. Thromb. Vasc. Biol.* **32**, 1824–1831 (2012).
22. Lv, W. et al. The drug target genes show higher evolutionary conservation than non-target genes. *Oncotarget* **7**, 4961–4971 (2016).
23. Motenko, H., Neuhauser, S. B., O'Keefe, M. & Richardson, J. E. MouseMine: a new data warehouse for MGI. *Mamm. Genome* **26**, 325–330 (2015).
24. Hart, T. et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 (Bethesda)* **7**, 2719–2727 (2017).
25. Finer, S. et al. Cohort profile: East London genes & health (ELGH), a community-based population genomics and health study of British Bangladeshi and British Pakistani people. *Int. J. Epidemiol.* **49**, 20–21 (2019).
26. Narasimhan, V. M. et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016).
27. Saleheen, D. et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).
28. Sulem, P. et al. Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).
29. DeBoever, C. et al. Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).
30. Tabrizi, S. J. et al. Targeting Huntingtin expression in patients with Huntington's disease. *N. Engl. J. Med.* **380**, 2307–2316 (2019).
31. DeVos, S. L. et al. Tau reduction prevents neuronal loss and reverses pathological tau deposition and seeding in mice with tauopathy. *Sci. Transl. Med.* **9**, eaag0481 (2017).
32. Minikel, E. V. et al. Prion protein lowering is a disease-modifying therapy across prion strains, disease stages, and endpoints. Preprint at https://doi.org/10.1101/2020.03.27.011940 (2020).
33. McCampbell, A. et al. Antisense oligonucleotides extend survival and reverse decrement in muscle response in ALS models. *J. Clin. Invest.* **128**, 3558–3567 (2018).
34. Chen, J., Chen, Y. & Pu, J. Leucine-rich repeat kinase 2 in Parkinson's disease: updated from pathogenesis to potential therapeutic target. *Eur. Neurol.* **79**, 256–265 (2018).
35. Cole, T. A. et al. Alpha-synuclein antisense oligonucleotides as a disease-modifying therapy for Parkinson's disease. Preprint at https://doi.org/10.1101/830554 (2019).
36. Minikel, E. V. et al. Age at onset in genetic prion disease and the design of preventive clinical trials. *Neurology* **93**, e125–e134 (2019).
37. Duyao, M. P. et al. Inactivation of the mouse Huntington's disease gene homolog Hdh. *Science* **269**, 407–410 (1995).
38. Rodan, L. H. et al. A novel neurodevelopmental disorder associated with compound heterozygous variants in the huntingtin gene. *Eur. J. Hum. Genet.* **24**, 1826–1827 (2016).
39. Ambrose, C. M. et al. Structure and expression of the Huntington's disease gene: evidence against simple inactivation due to an expanded CAG repeat. *Somat. Cell Mol. Genet.* **20**, 27–38 (1994).
40. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
41. Minikel, E. V. et al. Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9 (2016).
42. Zambrowicz, B. P. & Sands, A. T. Knockouts model the 100 best-selling drugs—will they model the next 100? *Nat. Rev. Drug Discov.* **2**, 38–51 (2003).
43. Uhl, K., Kennedy, D. L. & Kweder, S. L. Risk management strategies in the Physicians' Desk Reference product labels for pregnancy category X drugs. *Drug Saf.* **25**, 885–892 (2002).
44. Haggarty, S. J., Koeller, K. M., Wong, J. C., Grozinger, C. M. & Schreiber, S. L. Domain-selective small-molecule inhibitor of histone deacetylase 6 (HDAC6)-mediated tubulin deacetylation. *Proc. Natl Acad. Sci. USA* **100**, 4389–4394 (2003).
45. Zhang, B. W. et al. T cell responses in calcineurin A alpha-deficient mice. *J. Exp. Med.* **183**, 413–420 (1996).
46. Jacinto, E. et al. Mammalian TOR complex 2 controls the actin cytoskeleton and is rapamycin insensitive. *Nat. Cell Biol.* **6**, 1122–1128 (2004).
47. Hoshi, N., Langeberg, L. K., Gould, C. M., Newton, A. C. & Scott, J. D. Interaction with AKAP79 modifies the cellular pharmacology of PKC. *Mol. Cell* **37**, 541–550 (2010).
48. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
49. Whiffin, N. et al. Characterising the loss-of-function impact of 5′ untranslated region variants in whole genome sequence data from 15,708 individuals. *Nat. Commun.* https://doi.org/10.1038/s41467-019-10717-9 (2020).
50. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Genome Aggregation Database Production Team

Jessica Alföldi[1,4], Irina M. Armean[1,4,15], Eric Banks[16], Louis Bergelson[16], Kristian Cibulskis[16], Ryan L. Collins[1,17,18], Kristen M. Connolly[19], Miguel Covarrubias[16], Beryl B. Cummings[1,4,5], Mark J. Daly[1,4], Stacey Donnelly[1], Yossi Farjoun[16], Steven Ferriera[20], Laurent Francioli[1,4], Stacey Gabriel[20], Laura D. Gauthier[16], Jeff Gentry[16], Namrata Gupta[1,20], Thibault Jeandet[16], Diane Kaplan[16], Konrad J. Karczewski[1,4], Kristen M. Laricchia[1,4], Christopher Llanwarne[16], Eric V. Minikel[1,2,4], Ruchi Munshi[16], Benjamin M. Neale[1,4], Sam Novod[16], Anne H. O'Donnell-Luria[1,21,22], Nikelle Petrillo[16], Timothy Poterba[1,2,4], David Roazen[16], Valentin Ruano-Rubio[16], Andrea Saltzman[1], Kaitlin E. Samocha[9], Molly Schleicher[1], Cotton Seed[2,4], Matthew Solomonson[1,4], Jose Soto[16], Grace Tiao[1,4], Kathleen Tibbetts[16], Charlotte Tolonen[16], Christopher Vittal[2,4], Gordon Wade[16], Arcturus Wang[1,2,4], Qingbo Wang[1,4,18], James S. Ware[1,23,24], Nicholas A. Watts[1,4], Ben Weisburd[16] & Nicola Whiffin[1,23,24]

[15]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [16]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [17]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [18]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA. [19]Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [20]Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [21]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [22]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [23]National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. [24]Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London, UK.

Genome Aggregation Database Consortium

Carlos A. Aguilar Salinas[25], Tariq Ahmad[26], Christine M. Albert[27,28], Diego Ardissino[29], Gil Atzmon[30,31,32], John Barnard[33], Laurent Beaugerie[34], Emelia J. Benjamin[35,36,37], Michael Boehnke[38], Lori L. Bonnycastle[39], Erwin P. Bottinger[40], Donald W. Bowden[41,42,43], Matthew J. Bown[44], John C. Chambers[45,46,47], Juliana C. Chan[48], Daniel Chasman[27], Judy Cho[40], Mina K. Chung[33], Bruce Cohen[49,50], Adolfo Correa[51], Dana Dabelea[52], Mark J. Daly[1,4], Dawood Darbar[53], Ravindranath Duggirala[54], Josée Dupuis[55,56], Patrick T. Ellinor[1,57], Roberto Elosua[58,59,60], Jeanette Erdmann[61,62,63], Tõnu Esko[1,64], Martti Färkkilä[65], Jose Florez[1], Andre Franke[66], Gad Getz[67], Benjamin Glaser[68], Stephen J. Glatt[69], David Goldstein[70,71], Clicerio Gonzalez[72], Leif Groop[73], Christopher Haiman[74], Craig Hanis[75], Matthew Harms[76,77], Mikko Hiltunen[78], Matti M. Holi[79], Christina M. Hultman[80], Mikko Kallela[81], Jaakko Kaprio[82,83], Sekar Kathiresan[1,17,84], Bong-Jo Kim[85], Young Jin Kim[85], George Kirov[86], Jaspal Kooner[10,46,47],

Seppo Koskinen[87], Harlan M. Krumholz[88], Subra Kugathasan[89], Soo Heon Kwak[90], Markku Laakso[91,92], Terho Lehtimäki[93], Ruth J. F. Loos[40,94], Steven A. Lubitz[1,95], Ronald C. W. Ma[96,97,98], Daniel G. MacArthur[1,4,153,154], Jaume Marrugat[59,99], Kari M. Mattila[93], Steven McCarroll[2,100], Mark I. McCarthy[101,102,103], Dermot McGovern[104], Ruth McPherson[105], James B. Meigs[1,84,106], Olle Melander[107], Andres Metspalu[64], Benjamin M. Neale[1,4], Peter M. Nilsson[108], Michael C. O'Donovan[86], Dost Ongur[49,84], Lorena Orozco[109], Michael J. Owen[86], Colin N. A. Palmer[110], Aarno Palotie[1,4,82], Kyong Soo Park[90,111], Carlos Pato[112], Ann E. Pulver[113], Nazneen Rahman[114], Anne M. Remes[115], John D. Rioux[116,117], Samuli Ripatti[1,82,118], Dan M. Roden[119,120], Danish Saleheen[121,122,123], Veikko Salomaa[124], Nilesh J. Samani[44,125], Jeremiah Scharf[1,2,17], Heribert Schunkert[126,127], Moore B. Shoemaker[128], Pamela Sklar[129,130,131,155], Hilkka Soininen[132], Harry Sokol[34], Tim Spector[133], Patrick F. Sullivan[80,134], Jaana Suvisaari[124], E. Shyong Tai[135,136,137], Yik Ying Teo[135,138,139], Tuomi Tiinamaija[82,140,141], Ming Tsuang[142,143], Teresa Dan Turner[144], Teresa Tusie-Luna[145,146], Erkki Vartiainen[147], James. S. Ware[1,23,24], Hugh Watkins[148], Rinse K. Weersma[149], Maija Wessman[82,140], James G. Wilson[150] & Ramnik J. Xavier[151,152]

[25]Unidad de Investigacion de Enfermedades Metabolicas, Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico. [26]Peninsula College of Medicine and Dentistry, Exeter, UK. [27]Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [28]Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [29]Department of Cardiology, University Hospital, Parma, Italy. [30]Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel. [31]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, USA. [32]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. [33]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [34]Sorbonne Université, APHP, Gastroenterology Department, Saint Antoine Hospital, Paris, France. [35]Framingham Heart Study, National Heart, Lung, & Blood Institute and Boston University, Framingham, MA, USA. [36]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [37]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA. [38]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [39]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [40]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [41]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [42]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [43]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [44]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK. [45]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [46]Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK. [47]Imperial College Healthcare NHS Trust, Imperial College London, London, UK. [48]Department of Medicine and Therapeutics, Chinese University of Hong Kong, Hong Kong, Hong Kong. [49]Program for Neuropsychiatric Research, McLean Hospital, Belmont, MA, USA. [50]Department of Psychiatry, Harvard Medical School, Boston, MA, USA. [51]Department of Medicine, University of Mississippi Medical Center, Jackson, MI, USA. [52]Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA. [53]Department of Medicine and Pharmacology, University of Illinois at Chicago, Chicago, IL, USA. [54]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA. [55]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [56]Framingham Heart Study, National Heart, Lung, & Blood Institute and Boston University, Framingham, MA, USA. [57]Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [58]Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain. [59]Centro de Investigación Biomédica en Red Enfermedades Cardiovaculares (CIBERCV), Barcelona, Catalonia, Spain. [60]Departament of Medicine, Medical School, University of Vic-Central University of Catalonia, Vic, Catalonia, Spain. [61]Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany. [62]DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, Lübeck, Germany. [63]University Heart Center Lübeck, Lübeck, Germany. [64]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [65]Clinic of Gastroenterology, Helsinki University and Helsinki University Hospital, Helsinki, Finland. [66]Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany. [67]Cancer Genome Computational Analysis Group, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [68]Endocrinology and Metabolism Department, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [69]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA. [70]Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [71]Department of Genetics & Development, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [72]Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Cuernavaca, Mexico. [73]Genomics, Diabetes and Endocrinology, Lund University, Malmo, Sweden. [74]Lund University Diabetes Centre, Malmo, Sweden. [75]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA. [76]Department of Neurology, Columbia University, New York, NY, USA. [77]Institute of Genomic Medicine, Columbia University, New York, NY, USA. [78]Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland. [79]Department of Psychiatry, Helsinki University Central Hospital, Lapinlahdentie, Helsinki, Finland. [80]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [81]Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland. [82]Institute for Molecular Medicine FIMM, University of Helsinki, Helsinki, Finland. [83]Department of Public Health, University of Helsinki, Helsinki, Finland. [84]Department of Medicine, Harvard Medical School, Boston, MA, USA. [85]Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, South Korea. [86]MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Cardiff, UK. [87]Department of Health, THL-National Institute for Health and Welfare, Helsinki, Finland. [88]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA. [89]Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, GA, USA. [90]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. [91]Institute of Clinical Medicine, The University of Eastern Finland, Kuopio, Finland. [92]Kuopio University Hospital, Kuopio, Finland. [93]Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [94]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [95]Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA, USA. [96]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. [97]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [98]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. [99]Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain. [100]Department of Genetics, Harvard Medical School, Boston, MA, USA. [101]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Headington, Oxford, UK. [102]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [103]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [104]F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [105]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada. [106]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [107]Department of Clinical Sciences, University Hospital Malmo Clinical Research Center, Lund University, Malmo, Sweden. [108]Department of Clinical Sciences, Lund University, Skane University Hospital, Malmo, Sweden. [109]Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, Mexico. [110]Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. [111]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea. [112]Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, CA, USA. [113]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [114]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK. [115]Research Unit of Clinical Neuroscience, University of Oulu, Oulu, Finland. [116]Research Center, Montreal Heart Institute, Montreal, Quebec, Canada. [117]Department of Medicine, Faculty of Medicine, Université de Montréal, Quebec, Canada. [118]Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [119]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [120]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [121]Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [122]Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [123]Center for Non-Communicable Diseases, Karachi, Pakistan. [124]National Institute for Health and Welfare, Helsinki, Finland. [125]NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK. [126]Deutsches Herzzentrum München, Munich, Germany. [127]Technische Universität München, Munich, Germany. [128]Division of Cardiovascular Medicine, Nashville VA Medical Center and Vanderbilt University, School of Medicine, Nashville, TN, USA. [129]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [130]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [131]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [132]Institute of Clinical Medicine Neurology, University of Eastern Finland, Kuopio, Finland. [133]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [134]Department of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [135]Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore, Singapore. [136]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [137]Duke-NUS Graduate Medical School, Singapore, Singapore. [138]Life Sciences Institute, National University of Singapore, Singapore, Singapore. [139]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore. [140]Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland. [141]HUCH Abdominal Center, Helsinki University Hospital, Helsinki, Finland. [142]Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego, CA, USA. [143]Institute of Genomic Medicine, University of California, San Diego, CA, USA. [144]Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel. [145]Instituto de Investigaciones Biomédicas UNAM, Mexico City, Mexico. [146]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. [147]Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland. [148]Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [149]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. [150]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [151]Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [152]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. [155]Deceased: Pamela Sklar.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Data sources

pLoF analyses used the gnomAD dataset of 141,456 individuals[7]. For data consistency, all genome-wide constraint and CAF analyses used only the 125,748 gnomAD exomes. Curated analyses of individual genes used all 141,456 individuals including 15,708 whole genomes. Gene lists used in this study were extracted from public data sources between September 2018 and June 2019. Data sources and criteria for gene list extraction are shown in Extended Data Table 1. This study was performed under ethical approval from the Partners Healthcare Institutional Research Board (2013P001339/MGH) and the Broad Institute Office of Research Subjects Protection (ORSP-3862). All research participants provided informed consent.

Calculation of pLoF constraint

The calculation of constraint values for genes has been described in general elsewhere[10,12] and for this dataset specifically by Karczewski et al.[7]. Constraint calculations used LOFTEE-filtered ('high confidence') single-nucleotide variants (which for pLoF means nonsense and essential splice site mutations) found in gnomAD exomes with minor allele frequency <0.1%. Only unique canonical transcripts for protein-coding genes were considered, yielding 17,604 genes with available constraint values. For curated genes (Extended Data Table 2), the number of observed variants passing curation was divided by the expected number of variants to yield a curated constraint value. For *PRNP*, the expected number of variants was adjusted by multiplying by the ratio of the sum of mutation frequencies for all possible pLoF variants in codons 1–144 to the sum of mutation frequencies for all possible pLoF variants in the entire transcript, yielding 6 observed out of 6.06 expected. For *MAPT*, the expected number of variants was taken from Ensembl transcript ENST00000334239, which includes only the exons identified as constitutively brain-expressed in Fig. 3b (exon numbering previously described[51]).

Calculation of pLoF heterozygote and homozygote/compound heterozygote frequencies

LOFTEE-filtered high-confidence pLoF variants with minor allele frequency <5% in 125,748 gnomAD exomes were used to compute the proportion of individuals without a loss-of-function variant (q); the CAF was computed as $p = 1 - \mathrm{sqrt}(q)$. This approach conservatively assumes that, if an individual has two different pLoF variants, they are in *cis* to each other and count as only one pLoF allele.

For outbred populations (Fig. 2a), we used the value of p from all 125,748 gnomAD exomes, as this allows the largest possible sample size. This includes some individuals from bottlenecked populations, for which the distribution of p does differ from outbred populations, but these individuals are a small proportion of gnomAD exomes (12.6%). This also includes some consanguineous individuals, but these are an even smaller proportion of gnomAD exomes (2.3%), and any difference in the value of p between consanguineous and outbred populations is expected to be very small. Heterozygote frequency was calculated as $2p(1-p)$ and homozygote and compound heterozygote frequency was calculated as p^2. Lines indicate the size of gnomAD (141,456 individuals) and the world population (6.69 billion).

For bottlenecked populations (Fig. 2b), we used the value of p from the 10,824 Finnish exomes only. Lines indicate the number of Finnish individuals in gnomAD (12,526) and the population of Finland (5.5 million).

For consanguineous individuals (Fig. 2c), we again used the value of p from all gnomAD exomes, because p is not expected to differ greatly in consanguineous versus outbred populations. We used the mean proportion of the genome in runs of autozygosity (a) from individuals self-reporting second cousin or closer parents in East London Genes & Health, $a = 0.05766$ (rounded to 5.8%). Heterozygote frequency was calculated as $2p(1-p)$ and homozygote and compound heterozygote frequency was calculated as $(1 - a)p^2 + ap$. Lines indicate the number of consanguineous South Asian individuals in gnomAD ($n = 2{,}912$, by coincidence the same number as report second cousin or closer parents in ELGH) based on $F > 0.05$ (a conservative estimate, because second cousin parents are expected to yield $F = 0.015625$), and the estimated number of individuals in the world with second cousin or closer parents (10.4% of the world population)[9].

Several caveats apply to our CAF analysis. First, our approach naively treats genes with no pLoFs observed as having $P = 0$, even though pLoFs might be discovered at a larger sample size. Second, we naively group all populations together, even though the distribution of populations sampled in gnomAD does not reflect the world population[7]; we believe that this is reasonable because CAF for many genes is driven by singletons and other ultra-rare variants for which frequency is not expected to differ appreciably by continental population[10]. (It is important to note that the histograms shown in Fig. 2 reflect the expected frequency of heterozygotes and homozygotes/compound heterozygotes, based on gnomAD allele frequency, rather than the actual observed frequency of individuals with these genotypes in gnomAD.) Third, we use only protein-truncating variants annotated as pLoF in gnomAD. Structural and non-coding variation resulting in a loss of function may be missed in exomes, and missense variants resulting in a loss of function cannot be rigorously annotated. Fourth, we naively treat genes with one pLoF allele observed as having $P = 1/(2 \times 125{,}748)$, even though on average singleton variants have a true allele frequency lower than their nominal allele frequency[10]. Fifth, the variants included in this analysis are filtered but have not been manually curated or functionally validated, so some will ultimately prove not to be true LoF. These false positives tend to be more common and will have disproportionately contributed to the cumulative LoF allele frequency. Sixth, as described in the main text, our calculations assume that complete knockout is tolerated, which will not be true for some genes. We therefore also include a projection of the sample size needed to infer lethality from the absence of two-hit knockout individuals (Fig. 2e). Points one to three will tend to lead to underestimation of the true complete knockout frequency, whereas points four to six will tend to lead to overestimation. On balance, our calculations may reflect an upper bound of complete knockout frequency for most genes owing to the strong influence of factors five and six. Finally, as a matter of comparison between population structures, the sample size for all gnomAD exomes (Fig. 2a, c) is larger than for only Finnish exomes (Fig. 2b). For a version of Fig. 2 with the global gnomAD population downsampled to the same sample size as the gnomAD Finnish population, see Extended Data Fig. 2.

Knockout roadmap

For the knockout 'roadmap' (Fig. 2d, e), we classified genes according to the current status of human disease association and LoF ascertainment. Genes were classified as having a Mendelian disease association if they were present in OMIM with the filters described in Extended Data Table 1.

Remaining genes were classified as '2-hit LoF reported' based on presence in one or more of the following gene lists: homozygous LoF genotypes in gnomAD curated as previously described[7]; filtered homozygous LoF genotypes in runs of autozygosity with minor allele frequency <1% in canonical transcripts in the Bradford, Birmingham and ELGH[25] cohorts (total $n = 8{,}925$); observed number of imputed homozygotes >1 or number of compound heterozygous carriers where minor allele frequency <2% (for both variants) in deCODE[28]; homozygous LoF reported in PROMIS[27]; homozygous LoF with minor allele frequency <1% in UK Biobank[29].

The remainder of genes were sequentially classified as 'likely haploinsufficient' if pLI >0.9 in gnomAD, 'pLoF not yet observed' if CAF = 0 in gnomAD, and, finally, 'pLoF observed in gnomAD' if CAF >0 in gnomAD.

Genetic prevalence estimation

Here, we define 'genetic prevalence' for a given gene as the proportion of individuals in the general population at birth who have a pathogenic variant in that gene that will cause them to later develop disease. Genetic prevalence has not been well-studied or estimated for most disease genes.

In principle, it should be possible to estimate genetic prevalence simply by examining the allele frequency of reported pathogenic variants in gnomAD. In practice, three considerations usually preclude this approach. First, the present gnomAD sample size of 141,456 exomes and genomes is still too small to permit accurate estimates for very rare diseases. Second, the mean age of gnomAD individuals is approximately 55, which is above the age of onset for many rare genetic diseases, and individuals with known Mendelian disease are deliberately excluded, so pathogenic variants will be depleted in this sample relative to the whole birth population. Third and most importantly, a large fraction of reported pathogenic variants lack strong evidence for pathogenicity and are either benign or low penetrance[10,41], so without careful curation of pathogenicity assertions, summing the frequency of reported pathogenic variants in gnomAD will in most cases vastly overestimate the true genetic prevalence of a disease.

Instead, we searched the literature and very roughly estimated genetic prevalence based on available data. In most cases, we took disease incidence (new cases per year per population), multiplied by proportion of cases due to variants in a gene of interest, and multiplied by average age at death in cases. In some cases, estimates of at-risk population or direct measures of genetic prevalence were available. Details of the calculations undertaken for each gene are provided in Extended Data Table 4.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The gnomAD v2 data are available via the gnomAD browser (https://gnomad.broadinstitute.org).

Code availability

Additional data and the R 3.5.1 and Python 2.7.10 source code for this study are available via GitHub (https://github.com/ericminikel/drug_target_lof).

51. Andreadis, A. Tau splicing and the intricacies of dementia. *J. Cell. Physiol.* **227**, 1220–1225 (2012).
52. Yates, B. et al. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45** (D1), D619–D625 (2017).
53. Mainland, J. D., Li, Y. R., Zhou, T., Liu, W. L. L. & Matsunami, H. Human olfactory receptor responses to odorants. *Sci. Data* **2**, 150002 (2015).
54. Blekhman, R. et al. Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18**, 883–889 (2008).
55. Harding, S. D. et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Res.* **46** (D1), D1091–D1106 (2018).
56. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45** (D1), D896–D901 (2017).
57. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
58. Pringsheim, T. et al. The incidence and prevalence of Huntington's disease: a systematic review and meta-analysis. *Mov. Disord.* **27**, 1083–1091 (2012).
59. Keum, J. W. et al. The HTT CAG-expansion mutation determines age at death but not disease duration in huntington disease. *Am. J. Hum. Genet.* **98**, 287–298 (2016).
60. Kay, C. et al. Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology* **87**, 282–288 (2016).
61. Fisher, E. R. & Hayden, M. R. Multisource ascertainment of Huntington disease in Canada: prevalence and population at risk. *Mov. Disord.* **29**, 105–114 (2014).
62. Pringsheim, T., Jette, N., Frolkis, A. & Steeves, T. D. L. The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Mov. Disord.* **29**, 1583–1590 (2014).
63. Healy, D. G. et al. Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. *Lancet Neurol.* **7**, 583–590 (2008).
64. Onyike, C. U. & Diehl-Schmid, J. The epidemiology of frontotemporal dementia. *Int. Rev. Psychiatry* **25**, 130–137 (2013).
65. Bang, J., Spina, S. & Miller, B. L. Frontotemporal dementia. *Lancet* **386**, 1672–1682 (2015).
66. Maddox, R. A. et al. Prion disease incidence in the United States, 2003–2015. *Neurology* **94**, e153–e157 (2019).
67. Trinh, J., Guella, I. & Farrer, M. J. Disease penetrance of late-onset parkinsonism: a meta-analysis. *JAMA Neurol.* **71**, 1535–1539 (2014).
68. Chiò, A. et al. Prevalence of SOD1 mutations in the Italian ALS population. *Neurology* **70**, 533–537 (2008).
69. Cudkowicz, M. E. et al. Epidemiology of mutations in superoxide dismutase in amyotrophic lateral sclerosis. *Ann. Neurol.* **41**, 210–221 (1997).
70. Renton, A. E., Chiò, A. & Traynor, B. J. State of play in amyotrophic lateral sclerosis genetics. *Nat. Neurosci.* **17**, 17–23 (2014).
71. Byrne, S. et al. Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatry* **82**, 623–627 (2011).
72. Rowland, L. P. & Shneider, N. A. Amyotrophic lateral sclerosis. *N. Engl. J. Med.* **344**, 1688–1700 (2001).
73. Hirtz, D. et al. How common are the "common" neurologic disorders? *Neurology* **68**, 326–337 (2007).
74. Logroscino, G. et al. Incidence of amyotrophic lateral sclerosis in Europe. *J. Neurol. Neurosurg. Psychiatry* **81**, 385–390 (2010).
75. Hernandez, D. G., Reed, X. & Singleton, A. B. Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. *J. Neurochem.* **139** (Suppl. 1), 59–74 (2016).
76. Funayama, M. et al. A new locus for Parkinson's disease (PARK8) maps to chromosome 12p11.2-q13.1. *Ann. Neurol.* **51**, 296–301 (2002).
77. Zimprich, A. et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* **44**, 601–607 (2004).
78. Goldwurm, S. et al. Evaluation of LRRK2 G2019S penetrance: relevance for genetic counseling in Parkinson disease. *Neurology* **68**, 1141–1143 (2007).
79. Do, C. B. et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet.* **7**, e1002141 (2011).
80. Kinoshita, T. & Fujita, M. Biosynthesis of GPI-anchored proteins: special emphasis on GPI lipid remodeling. *J. Lipid Res.* **57**, 6–24 (2016).
81. Kitamoto, T., Iizuka, R. & Tateishi, J. An amber mutation of prion protein in Gerstmann–Sträussler syndrome with mutant PrP plaques. *Biochem. Biophys. Res. Commun.* **192**, 525–531 (1993).
82. Finckh, U. et al. High prevalence of pathogenic mutations in patients with early-onset dementia detected by sequence analyses of four different genes. *Am. J. Hum. Genet.* **66**, 110–117 (2000).
83. Jayadev, S. et al. Familial prion disease with Alzheimer disease-like tau pathology and clinical phenotype. *Ann. Neurol.* **69**, 712–720 (2011).
84. Fong, J. C. et al. Genetic prion disease caused by PRNP Q160X mutation presenting with an orbitofrontal syndrome, cyclic diarrhea, and peripheral neuropathy. *J. Alzheimers Dis. JAD* **55**, 249–258 (2017).
85. Bommarito, G. et al. A novel prion protein gene-truncating mutation causing autonomic neuropathy and diarrhea. *Eur. J. Neurol.* **25**, e91–e92 (2018).
86. Mead, S. et al. A novel prion disease associated with diarrhea and autonomic neuropathy. *N. Engl. J. Med.* **369**, 1904–1914 (2013).
87. Capellari, S. et al. Two novel *PRNP* truncating mutations broaden the spectrum of prion amyloidosis. *Ann. Clin. Transl. Neurol.* **5**, 777–783 (2018).
88. Matsuzono, K. et al. A novel familial prion disease causing pan-autonomic-sensory neuropathy and cognitive impairment. *Eur. J. Neurol.* **20**, e67–e69 (2013).
89. Jansen, C. et al. Prion protein amyloidosis with divergent phenotype associated with two novel nonsense mutations in PRNP. *Acta Neuropathol.* **119**, 189–197 (2010).

Acknowledgements This study was performed under ethical approval from the Partners Healthcare Institutional Research Board (2013P001339/MGH) and the Broad Institute Office of Research Subjects Protection (ORSP-3862) in compliance with all relevant ethical regulations; written informed consent was obtained from all research participants. We thank all of the research participants for contributing their data. E.V.M. acknowledges support from the National Institutes of Health (NIH) (F31 AI122592) and an anonymous organization. gnomAD data aggregation was supported primarily by the Broad Institute, gnomAD analysis was supported in part by NIDDK U54 DK105566, and development of LOFTEE by NIGMS R01 GM104371. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. E.L.G.H. is funded by the Wellcome Trust (102627, 210561), the Medical Research Council (M009017), Higher Education Funding Council for England Catalyst, Barts Charity (845/1796), Health Data Research UK (for London substantive site), and research delivery support from the NHS National Institute for Health Research Clinical Research Network (North Thames). N.W. is supported by a Rosetrees and Stoneygate Imperial College Research Fellowship. The results published here are in part based upon data: (1) generated by The Cancer Genome Atlas managed by the NCI and NHGRI (accession: phs000178.v10.p8). Information about TCGA can be found at http://cancergenome.nih.gov; (2) generated by the Genotype-Tissue Expression Project (GTEx) managed by the NIH Common Fund and NHGRI (accession: phs000424.v7.p2); (3) generated by the Exome Sequencing Project, managed by NHLBI; (4) generated by the Alzheimer's Disease Sequencing Project (ADSP), managed by the NIA and NHGRI (accession: phs000572.v7.p4). We thank J. Kaprio and M. Kurki (Finnish Twins AD cohort) and Academy of Finland grant 312073, and Ruth McPherson

(Ottawa Genomics Heart Study) for providing information on individuals with *PRNP*-truncating variants. We thank J. B. Carroll, K. Heilbron, J. Fah Sathirapongsasuti, and L. C. Francioli for comments and suggestions. A subset of the analyses reported here originally appeared as a blog post on CureFFI.org (http://www.cureffi.org/2018/09/12/lof-and-drug-safety/).

Author contributions Conceived and designed the study: E.V.M., S.L.S., D.G.M. Performed analysis: E.V.M., K.J.K., H.C.M., B.B.C., N.W., D.R. Supervised the research: J.A., R.C.T., D.A.v.H., M.J.D., S.L.S., D.G.M. Provided data: gnomAD consortium (Genome Aggregation Database Production Team and Genome Aggregation Database Consortium), H.C.M., R.C.T., D.A.v.H. Wrote the paper: E.V.M. Edited and approved the final manuscript: all authors

Competing interests E.V.M. has received research support in the form of charitable contributions from Charles River Laboratories and Ionis Pharmaceuticals, and has consulted for Deerfield Management. K.J.K. is a shareholder of Personalis. H.C.M., B.B.C., M.W., D.R. and J.A. have no competing interests to declare. R.C.T. serves on the Scientific Advisory Board of Ipsen Ltd and has current funding from the Wellcome Trust and the National Institute for Health Research UK. D.A.v.H. is a shareholder of Nexpep Pty Ltd; has current or recent research funding from Wellcome Trust, Medical Research Council UK, National Institute for Health Research UK, Alnylam Pharmaceuticals; and serves on the Population & Systems Medicine Board of the Medical Research Council UK. MJD is a founder of Maze Therapeutics. S.L.S. serves on the Board of Directors of the Genomics Institute of the Novartis Research Foundation ('GNF'); is a shareholder and serves on the Board of Directors of Jnana Therapeutics; is a shareholder of Forma Therapeutics; is a shareholder and advises Decibel Therapeutics and Eikonizo Therapeutics; serves on the Scientific Advisory Boards of Eisai Co., Ltd., Ono Pharma Foundation, Exo Therapeutics, and F-Prime Capital Partners; and is a Novartis Faculty Scholar. D.G.M. is a founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme.

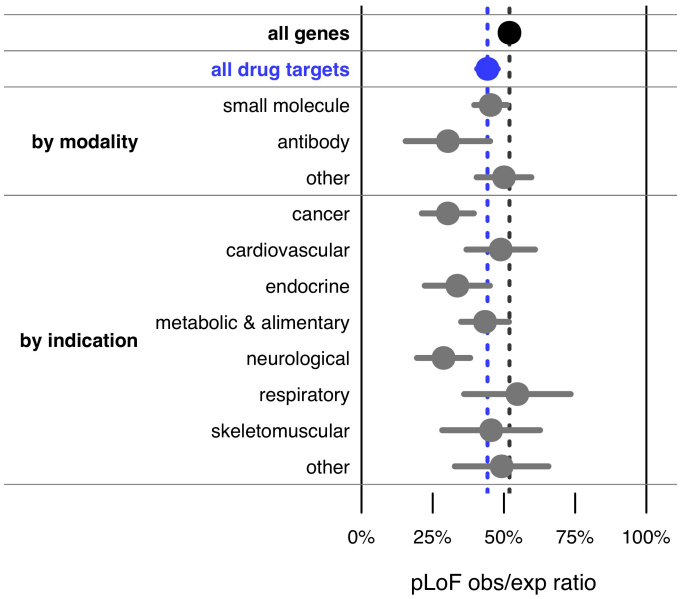
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2267-z.

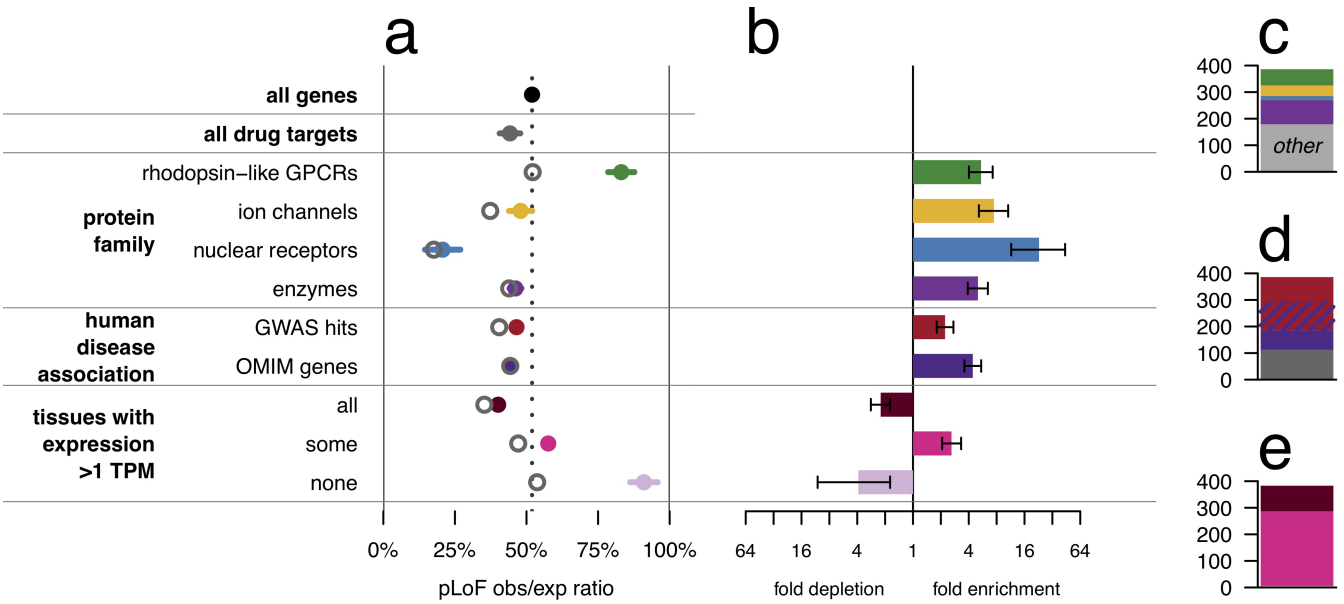
Correspondence and requests for materials should be addressed to E.V.M. or D.G.M.

Peer review information *Nature* thanks Paul de Bakker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

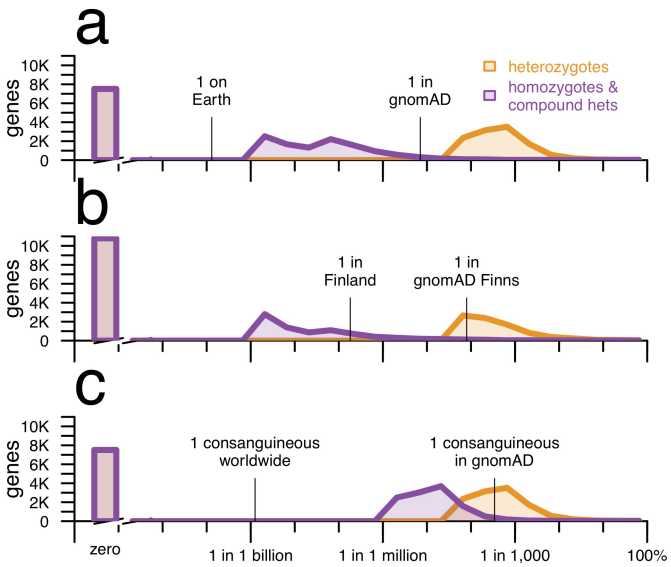
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Drug target constraint by modality and indication. Mean (dots) and 95% confidence interval (line segments) for constraint in subsets of drug-targets sets (data sources and number of genes for each list are provided in Extended Data Table 1). Modality information was extracted from DrugBank and indication information from ATC codes; see Extended Data Table 1 for details.



Extended Data Fig. 2 | Drug-target gene set confounding. a, Forest plot of means (dots) and 95% confidence intervals of the mean (line segments) for gene sets evaluated for confounding with drug-target status. Data sources and number of genes for each list are provided in Extended Data Table 1. LoF obs/exp ratios differ significantly from the set of all genes for four canonically druggable protein families (top), human disease-associated genes (middle), and genes by broadness of tissue expression (bottom). Within each class, the genes that are drug targets have a lower mean obs/exp ratio (hollow grey circles) than the class overall. **b**, The druggable protein families, disease-associated genes, and genes expressed in some tissues but not others are enriched several-fold among the set of drug targets. Bars indicate fold enrichment and error bars indicate 95% confidence intervals. **c**–**e**, Composition of drug targets when broken down by protein family (**c**), disease association (**d**), or broadness of tissue expression (**e**). The enriched classes account for most drug targets. In a linear model, after controlling for protein family, disease association status, and number of tissues with expression >1 transcript per million (TPM), drug targets are still more constrained than other genes (−8.0% obs/exp, nominal $P = 0.00011$, $t = -3.9$, df = 17,325 for the contribution of drug_target in the linear regression obs/exp ~ drug_target + family + dz_assoc + n_tissues), but the probable existence of additional unobserved confounders cautions against over-interpretation of this observation (see main text).



Extended Data Fig. 3 | Expected frequency of individuals with one or two null alleles for every protein-coding gene across different population models, with sample size held constant. This is identical to Fig. 2 except as follows. As noted in the Methods, one caveat about Fig. 2 is that the sample size is larger for the plots using all gnomAD exomes (Fig. 2a, c) than for Finnish exomes (Fig. 2b). This figure shows the same analysis, but with the global gnomAD population downsampled to 10,824 randomly chosen exomes so that the same size is identical to that of Finnish exomes. Computation of $P = 1 - \mathrm{sqrt}(q)$ as described in the Methods is computationally expensive for downsampled datasets because it requires individual-level genotypes. Instead, this analysis uses 'classic' CAF, which is simply the sum of allele frequencies of all high-confidence pLoF variants each at allele frequency <5%, capped at a total of 100%, for both global and Finnish exomes. The results show that even when the sample size is held constant, the number of genes with zero pLoF variants observed is higher in a bottlenecked population than in a mostly outbred population. A constant *y* axis with no axis breaks is used in this figure to make this difference more clearly visible.

Extended Data Table 1 | Data sources for gene lists used in this study

list	*N*	*N**	reference / criteria
All	19,194	17,604	HGNC[52].
Olfactory receptors	371	325	Mainland et al[53].
Homozygous LoF tolerant	330	325	≥2 different high-confidence pLoF variants each homozygous in ≥1 individual in gnomAD exomes.
Autosomal recessive	527	519	Blekhman et al[54].
Autosomal dominant	307	305	Blekhman et al[54].
Essential in culture	683	659	Hart et al[24].
ClinGen haploinsufficient	294	288	ClinGen Dosage Sensitivity Map[18] level 3
Approved drug targets	386	383	DrugBank 5.0 XML release[17] (acc. Sep 12, 2018); Top-ranked mechanistic target of approved drugs. group==’approved’, target.attrib[‘position’] == '1', known-action==’yes’
Positive targets	143	142	DrugBank action: activator, agonist, chaperone, cofactor, gene replacement, inducer, partial agonist, positive allosteric modulator, positive modulator, potentiator, stimulator
Negative targets	243	241	DrugBank action: antagonist, blocker, degradation, inhibitor, inverse agonist, negative modulator, neutralizer, suppressor
Other & unknown (effect)	94	94	DrugBank action other or unlisted.
Small molecule	176	175	DrugBank type == ‘small’
Antibody	18	18	DrugBank type == ‘biotech’ and ‘Antibodies’ in categories
Other (modality)	35	35	DrugBank type == ‘biotech’ and ‘Antibodies’ not in categories
Oncology	45	45	ATC level 1 code L
Cardiovascular	38	38	ATC level 1 code C
Endocrine	24	24	ATC level 1 code G or H
Metabolic & alimentary	38	38	ATC level 1 code A
Neurology	35	35	ATC level 1 code N
Respiratory	12	11	ATC level 1 code R
Skeletomuscular	14	14	ATC level 1 code M
Other (indication)	29	28	ATC level 1 code B, D, J, P, S, or V
Rhodopsin-like GPCRs	689	604	HGNC[52] gene set 140: “G protein-coupled receptors, Class A rhodopsin-like”.
Ion channels	326	323	HGNC gene set 177: “Ion channels”[52].
Nuclear receptors	48	47	IUPHAR/BPS Guide to Pharmacology “Nuclear receptors”[55] .
Enzymes	1,178	1,144	IUPHAR/BPS Guide to Pharmacology “Enzymes”[55].
GWAS hits	6,336	6,080	GWAS Catalog[56] MAPPED_GENE column (P < 5-e8)
OMIM genes	3,367	3,294	OMIM[57] (acc. June 11, 2019) phenotypes with MIM number, lacking ‘?’, ‘{‘, ‘[‘, “response”, “susceptibility”, or “somatic”.
All (tissues)	7,931	7,550	>1 TPM in all 53 tissues in GTEx[40] v7
Some (tissues)	9,698	9,009	>1 TPM in >0 and <53 tissues in GTEx[40] v7
None (tissues)	1,076	776	>1 TPM in 0 tissues in GTEx[40] v7
Mouse heterozygous lethal knockout	401	395	MouseMine[23]

For analysis, only protein-coding genes with unambiguous mapping to current approved gene symbols were used; numbers in the table reflect this. Values in the *N* column indicate totals from the full universe of 19,194 genes; values in the *N** column indicate the subset of genes with non-missing constraint values, used for Fig. 1 and Extended Data Figs. 1, 2. The following references are cited in the table: refs. [17,18,23,24,40,52–57].

Extended Data Table 2 | Spectrum of tolerance to genetic inactivation among human drug targets

drug class	example	gene	obs/exp pLoF
topoisomerase I inhibitors	irinotecan	*TOP1*	0% (0/50.5)
M1-selective antimuscarinics	pirenzepine	*CHRM1*	0% (0/14.1)
cytoskeleton disruptors	paclitaxel	*TUBB*	6% (1/16.4)
non-steroidal anti-inflammatory drugs	aspirin	*PTGS2*	10% (3/29.7)
statins	atorvastatin	*HMGCR*	13% (6/46.3)
phosphodiesterase 5 inhibitors	sildenafil	*PDE5A*	33% (16/47.8)
antifolates	methotrexate	*DHFR*	38% (4/10.5)
proton pump inhibitors	omeprazole	*ATP4A*	52% (25/47.9)
antiplatelets	clopidogrel	*P2RY12*	66% (5/7.6)
H1 antihistamines	cetirizine	*HRH1*	76% (11/14.5)
angiotensin converting enzyme inhibitors	benazepril	*ACE*	87% (62/71.3)
cholesterol-lowering antibodies	alirocumab	*PCSK9*	98% (26/26.5)

Example targets are arranged from the most intolerant (top) to the most tolerant (bottom) of inactivation.

Extended Data Table 3 | Curation of pLoF variation in six neurodegenerative disease genes

gene	length (bp)	pLoF obs/exp	cumulative pLoF allele frequency		pLoF heterozygote frequency	GoF disease genetic prevalence
			before filtering & curation	after filtering & curation		
HTT	9,426	8.2%	6.2%	0.013%	1 in 3,800	1 in 2,400-4,400
LRRK2	7,581	41%	0.23%	0.09%	1 in 500	1 in 3,300
MAPT	2,328	0%[a]	14%	0%	—	1 in 5,000 – 31,000
PRNP	759	99%[b]	0.0035%	0.0021%	1 in 18,000	1 in 50,000
SNCA	420	0%	0.0012%	0%	—	1 in 360,000
SOD1	462	18%	0.0060%	0.0038%	1 in 26,000	1 in 27,000-83,000

Shown are the coding sequence length (base pairs, bp), constraint value (pLoF obs/exp) after filtering and curation, cumulative allele frequency before and after filtering and manual curation, estimated frequency of true pLoF heterozygotes in the population, and genetic prevalence (population frequency including pre-symptomatic individuals) of the GoF disease associated with the gene. Genetic prevalence calculations are described in Extended Data Table 4, and variant curation details are provided in Supplementary Table 1, except for *LRRK2*, which is described in detail elsewhere[49].

[a]Constitutive brain-expressed exons only.

[b]*PRNP* codons 1–144; see Fig. 3c for rationale.

Extended Data Table 4 | Estimation of genetic prevalence for GoF genetic neurodegenerative diseases

gene	basis for genetic prevalence estimation	estimate
HTT	A reported HD incidence of 0.38 cases per 100,000 per year based on meta-analysis[58] multiplied by an average age at death of ~60 for the most common CAG lengths[59]. Finally, a genetic screen of a general population sample[60] found ≥40 CAG repeat alleles, which are presumed to be fully penetrant, in 3 individuals out of 7,315, for a genetic prevalence of 1 in 2,438.	1 in 4,386
HTT	Prevalence of 13.7 per 100,000 symptomatic plus 81.6 per 100,000 at 25-50% risk in an exhaustive ascertainment study[61]. Assuming there are twice as many individuals at 25% risk as at 50% risk, then on average 33.3% of the 81.6, or 27.1 per 100,000 have the mutation. Thus, 13.7 + 27.1 = 40.8 per 100,000 individuals have an *HTT* CAG expansion.	1 in 2,451
HTT	A genetic screen of a general population sample[60] found ≥40 CAG repeat alleles, which are presumed to be fully penetrant, in 3 individuals out of 7,315.	1 in 2,438
LRRK2	Based on meta-analysis[62], Parkinson's disease (PD) has an estimated prevalence of 1,903 per 100,000 at age ≥80, meaning the general population's lifetime risk of PD is ~1.9%. It is generally stated that about 10% of PD cases are "familial" and the remainder sporadic; in a diverse worldwide case series, *LRRK2* mutations were found in 179/14,253 (1.3%) sporadic cases and 201/5,123 (3.9%) familial cases[63], implying that *LRRK2* mutations are present in ~1.6% of all PD cases. Thus, *LRRK2* mutations account for a 1.6% * 1.9% = ~0.030% lifetime risk of PD in the general population[a].	1 in 3,300
MAPT	Pathogenic *MAPT* mutations can present with a variety of clinical phenotypes, and common *MAPT* haplotypes are associated with risk for a variety of different neurodegenerative disorders; we were unable to identify any studies of genetic prevalence nor any large case series for any *MAPT*-associated phenotype. As a crude estimate, frontotemporal dementia has a reported incidence of 2.7-4.1 per 100,000 per year[64] with typical age at death of perhaps 60, and *MAPT* mutations accounting for 5-20% of familial cases, and familial cases accounting for 40% of all cases[65]. Multiplying all these figures results in range of 0.0032% to 0.020%.	1 in 5,000 – 31,000
PRNP	We recently considered the lifetime risk of genetic prion disease in detail[36]. Prion disease (including sporadic, genetic, and acquired) causes ~1 in 5,000 people based on either death certificate analysis or division of disease incidence by the overall death rate[41,66]. ~10% of cases are attributable to *PRNP* variants with evidence for Mendelian segregation (although additional cases harbor lower-penetrance variants)[36]. Thus, we expect a genetic prevalence of 1 in 50,000. On the order of ~1 in 100,000 people in gnomAD and 23andMe harbor high-penetrance *PRNP* variants[36,41], although as noted above, we expect these datasets to be depleted compared to the population at birth, because prion disease is rapidly fatal and many individuals in these databases are above the typical age of onset.	1 in 50,000
SNCA	As explained above for *LRRK2*, we assumed a 1.9% lifetime risk of Parkinson's disease (PD) in the general population, with 10% of cases being familial. *SNCA* point mutations, duplications, and triplications all appear to be highly penetrant, and in a familial PD case series these accounted for 103/709 = 15% of individuals[67]. Thus, we estimate that *SNCA* mutations account for a 1.9% * 10% * 15% = 0.00028% risk of PD in the general population.	1 in 360,000
SOD1	*SOD1* mutations are believed to account for ~12% to 24% of familial ALS[68,69] and 1% of sporadic ALS[68,70]. One a meta-analysis found that ~4.6% of ALS is familial[71], although a figure of 10% is also often used[72]. These figures imply that ~1.5 – 3.3% of all ALS is attributable to *SOD1*. The overall incidence of ALS is reported at ~1.6 – 2.2 per 100,000 per year[73,74], so the incidence of *SOD1* ALS might be estimated at ~0.024 – 0.073 per 100,000 per year. Age at death of ~50 is around average for many *SOD1* mutations[69], implying a 1.2 – 3.7 per 100,000 population prevalence of pathogenic *SOD1* mutations.	1 in 27,000-83,000

Data sources were identified by PubMed and Google Scholar searches. Genetic prevalence was defined as the proportion of the population at birth carrying a mutation and destined to later develop disease, and estimated as described for each gene. The following references are cited in the table: refs. [36,41,58–74].

[a]It is important to consider how this figure relates to the penetrance of *LRRK2* mutations, as *LRRK2* variants appear to occupy a spectrum of penetrance[75]. Some variants exhibit Mendelian segregation with disease[76,77], implying high risk; the G2019S variant is estimated to have approximately 32% penetrance[78]; and other common variants are risk factors with odds ratios of only around 1.2 estimated through genome-wide association studies (GWAS)[79]. The GWAS-implicated common variants were not included in the case series on which our estimate is based[63], but G2019S does account for most cases in that series. Because the 0.03% estimate here is based on counting symptomatic cases rather than asymptomatic individuals, it will appropriately underestimate the number of G2019S carriers. In essence, in this calculation each G2019S carrier in the population only counts as 1/3 of a person, because they have only a 1/3 probability of developing a disease. It is therefore appropriate that our estimate of genetic prevalence (0.03%) is actually lower than double the allele frequency of G2019S in gnomAD (0.1%).

Extended Data Table 5 | Details of *PRNP*-truncating variants

variant	allele count	neurological phenotype	comments	reference
G20Gfs84X	1	healthy	As previously reported.	[41]
R37X	2	healthy, unknown	One previously reported, one new.	[41], this work
Q41X	1	unknown		this work
H69 frameshifts	2	N/A	False variant calls in gnomAD, apparent alignment artifact due to octapeptide repeat region.	this work
Q75X	1	healthy	As previously reported	[41]
W81X	1	unknown		this work
W99X	1	unknown		this work
G131X	1	healthy	The presence of this variant in the ExAC database was previously reported, but without phenotype information. We now report that this individual is a 77-year-old male, cognitively well with no family history of dementia. Ascertained as a case in a study of coronary artery disease, this individual has hypertension and well-controlled dyslipidemia and has undergone one bypass surgery. He has two adult children.	[41], this work
Y145X	1	dementia		[81]
Q160X	5	dementia		[82–84]
Y162X	1	dementia		[85]
Y163X	7	dementia		[86,87]
Y169X	2	dementia		[87]
D178Efs25X	1	dementia		[88]
Q186X	1	dementia		[41]
Y226X	1	dementia		[89]
Q227X	1	dementia		[89]
L234Pfs7X	1	dementia	Ascertained as a female case in the Finnish twins Alzheimer disease cohort. Died at age >90 of proximal cause pneumonia, ultimate cause diagnosed as Alzheimer disease based on clinical examination only. Had a dizygotic twin not included in gnomAD.	this work

Allele count for variants from the literature in Fig. 3c is the total number of definite or probable cases with sequencing performed in the studies cited in this table. The L234Pfs7X variant changes the C-terminal GPI signal of prion protein from SMVLFSSPPVILLISFLIFLIVGX to SMVPSPLHLX. This new sequence does not adhere to the known rules of GPI anchor attachment[80]: GPI signals must contain a 5–10-polar-residue spacer followed by 15–20 hydrophobic residues. Thus, this frameshifted prion protein would be predicted to be secreted and thus may be pathogenic, explaining the Alzheimer's disease diagnosis in this individual. However, it is also possible that the new C-terminal sequence found here interferes with prion formation, and/or that this variant is incompletely penetrant, and that the diagnosis of Alzheimer's disease in this individual is merely a coincidence. The following references are cited in the table: refs. [41,81–89].

nature research

Corresponding author(s): Eric Vallabh Minikel and Daniel G. MacArthur

Last updated by author(s): 11/13/2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Analyses utilized Python 2.7.10 and R 3.5.1. Data and code sufficient to produce the plots and analyses in this paper are available at https://github.com/ericminikel/drug_target_lof
Data analysis	Analyses utilized Python 2.7.10 and R 3.5.1. Data and code sufficient to produce the plots and analyses in this paper are available at https://github.com/ericminikel/drug_target_lof

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Analyses utilized Python 2.7.10 and R 3.5.1. Data and code sufficient to produce the plots and analyses in this paper are available at https://github.com/ericminikel/drug_target_lof

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study was opportunistic, and involved secondary use of all available genome and exome data. No sample size was predetermined. Our flagship analysis of gnomAD loss-of-function variants (Karczewski et al, https://doi.org/10.1101/531210) indicates that the dataset is well-powered to examine constraint against such variants — for instance, 72% of genes have at least 10 pLoF variants expected in this sample size based on mutation rates.
Data exclusions	Sample QC and variant QC for the gnomAD database are described extensively by Karczewski et al, https://doi.org/10.1101/531210. Notably, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded.
Replication	We did not attempt to reproduce any findings in a separate dataset, as no other exome or genome sequencing dataset of comparable size exists.
Randomization	As this was a population-based study, and not a case-control study, no randomization was performed.
Blinding	As this was a population-based study, and not a case-control study, blinding was not relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	As an opportunistic collection of data, the participants in gnomAD were not selected based on age, gender, or genotypic information. As described above, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded. The population and dataset inclusion criteria are described in more detail by Karczewski et al, https://doi.org/10.1101/531210
Recruitment	The generation of the gnomAD database was an opportunistic secondary use study, we did not recruit any participants. The study is described in more detail by Karczewski et al, https://doi.org/10.1101/531210
Ethics oversight	This study was performed under ethical approval from the Partners Healthcare Institutional Research Board (2013P001339/MGH) and the Broad Institute Office of Research Subjects Protection (ORSP-3862) in compliance with all relevant ethical regulations; informed consent was obtained from all research participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Virological assessment of hospitalized patients with COVID-2019

https://doi.org/10.1038/s41586-020-2196-x

Received: 1 March 2020

Accepted: 24 March 2020

Published online: 1 April 2020

Check for updates

Roman Wölfel[1,6], Victor M. Corman[2,6], Wolfgang Guggemos[3,6], Michael Seilmaier[3], Sabine Zange[1], Marcel A. Müller[2], Daniela Niemeyer[2], Terry C. Jones[2,4], Patrick Vollmar[1], Camilla Rothe[5], Michael Hoelscher[5], Tobias Bleicker[2], Sebastian Brünink[2], Julia Schneider[2], Rosina Ehmann[1], Katrin Zwirglmaier[1], Christian Drosten[2,7]✉ & Clemens Wendtner[3,7]✉

Coronavirus disease 2019 (COVID-19) is an acute infection of the respiratory tract that emerged in late 2019[1,2]. Initial outbreaks in China involved 13.8% of cases with severe courses, and 6.1% of cases with critical courses[3]. This severe presentation may result from the virus using a virus receptor that is expressed predominantly in the lung[2,4]; the same receptor tropism is thought to have determined the pathogenicity—but also aided in the control—of severe acute respiratory syndrome (SARS) in 2003[5]. However, there are reports of cases of COVID-19 in which the patient shows mild upper respiratory tract symptoms, which suggests the potential for pre- or oligosymptomatic transmission[6–8]. There is an urgent need for information on virus replication, immunity and infectivity in specific sites of the body. Here we report a detailed virological analysis of nine cases of COVID-19 that provides proof of active virus replication in tissues of the upper respiratory tract. Pharyngeal virus shedding was very high during the first week of symptoms, with a peak at 7.11×10^8 RNA copies per throat swab on day 4. Infectious virus was readily isolated from samples derived from the throat or lung, but not from stool samples—in spite of high concentrations of virus RNA. Blood and urine samples never yielded virus. Active replication in the throat was confirmed by the presence of viral replicative RNA intermediates in the throat samples. We consistently detected sequence-distinct virus populations in throat and lung samples from one patient, proving independent replication. The shedding of viral RNA from sputum outlasted the end of symptoms. Seroconversion occurred after 7 days in 50% of patients (and by day 14 in all patients), but was not followed by a rapid decline in viral load. COVID-19 can present as a mild illness of the upper respiratory tract. The confirmation of active virus replication in the upper respiratory tract has implications for the containment of COVID-19.

There is a close genetic relationship between SARS coronavirus (SARS-CoV) and the causative agent of COVID-19, SARS-CoV-2. The predominant expression of ACE2 in the lower respiratory tract is believed to have determined the natural history of SARS as an infection of the lower respiratory tract[5]. Although the positive detection of SARS-CoV-2 in clinical specimens from the upper respiratory tract has previously been described[9,10], these observations do not address the principal differences between SARS and COVID-19 in terms of clinical pathology. The patients who were studied here were enrolled because they acquired their infections upon known close contact to an index case, thereby avoiding representational biases owing to symptom-based case definitions. All patients were treated in a single hospital in Munich, Germany. Virological testing was done by two closely collaborating laboratories that used the same standards of technology for PCR with reverse transcription (RT–PCR) and virus isolation; these two laboratories confirmed each other's results in almost all of the individual samples. Owing to the extremely high congruence of results, all data—except for the serological data (which are based on results from one laboratory only)—are presented together. The patients are part of a larger cluster of epidemiologically linked cases that occurred after 23 January 2020 in Munich, as discovered on 27 January (ref. [11]). The present study uses samples taken during the clinical course in the hospital, as well as from initial diagnostic testing before admission. In cases in which this initial diagnostic testing was done by other laboratories, the original samples were retrieved and retested under the rigorous quality standards of the present study.

RT–PCR, replication sites and infectivity

To first understand whether the described clinical presentations are solely caused by infection with SARS-CoV-2, samples from all patients

[1]Bundeswehr Institute of Microbiology, Munich, Germany. [2]Charité Universitätsmedizin Berlin, Berlin, Germany. [3]Klinikum München-Schwabing, Munich, Germany. [4]Center for Pathogen Evolution, Department of Zoology, University of Cambridge, Cambridge, UK. [5]University Hospital LMU Munich, Munich, Germany. [6]These authors contributed equally: Roman Wölfel, Victor M. Corman, Wolfgang Guggemos. [7]These authors jointly supervised this work: Christian Drosten, Clemens Wendtner. ✉e-mail: christian.drosten@charite.de; clemens.wendtner@muenchen-klinik.de

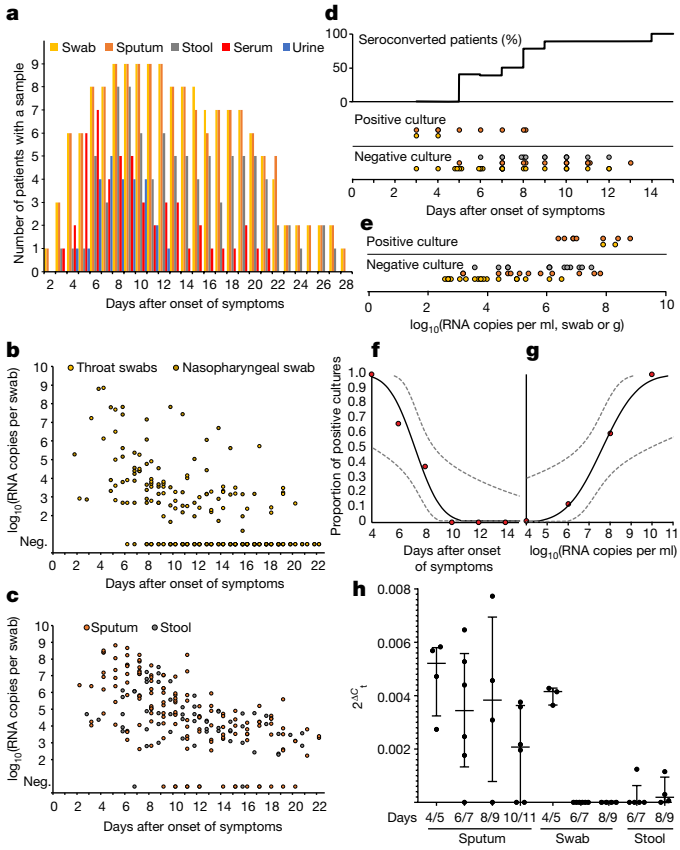


Fig. 1 | Hallmarks of viral shedding in aggregated samples. a, Samples and sample types per day. **b**, Viral RNA concentrations in samples from the upper respiratory tract. Neg., sample negative for RNA copies. **c**, Viral RNA concentrations in sputum and stool samples. **d**, Seroconversion and virus isolation success, dependent on day after the onset of symptoms. Top, fraction of seroconverted patients. Bottom, aggregated results of virus isolation trials. **e**, Virus isolation success, dependent on viral load. Viral loads were projected to RNA copies per ml (for sputum samples), per swab (for throat swab samples) or per g (for stool samples). **f**, **g**, Projected virus isolation success based on probit distributions. The inner lines are probit curves (dose–response rule). The outer dotted lines are 95% confidence interval. For a <5% isolation success, the estimated day was 9.78 (95% confidence interval 8.45–21.78) days after the onset of symptoms, and the estimated RNA concentration for <5% isolation success was estimated to be 5.40 $\log_{10}$(RNA copies per ml) (95% confidence interval −4.11–6.51). **h**, Subgenomic viral RNA transcripts in relation to viral genomic RNA. Dots represent mean values of RT–PCR data obtained from at least two independent experiments on samples from individual patients. Plots show median values with interquartile ranges.

were tested against a panel of typical agents of respiratory viral infection, including human coronavirus (HCoV)-HKU1, HCoV-OC43, HCoV-NL63 and HCoV-229E, influenza virus A, influenza virus B, phinovirus, enterovirus, respiratory syncytial virus, human parainfluenza viruses 1–4, human metapneumovirus, adenovirus and human bocavirus. No coinfection was detected in any patient.

All patients were initially diagnosed by RT–PCR from oro- or nasopharyngeal swab specimens[12]. Both types of specimen were collected over the whole clinical course in all patients. There were no discernible differences in viral loads or detection rates when comparing naso- and oropharyngeal swabs (Fig. 1b). The earliest swabs were taken on day 1 of symptoms, which were often very mild or prodromal. All swabs from all patients taken between day 1 and day 5 tested positive. The average virus RNA load was 6.76×10^5 copies per whole swab until day 5, and the maximum load was 7.11×10^8 copies per swab. Swab samples taken after day 5 had an average viral load of 3.44×10^5 copies per swab and

Table 1 | Single-nucleotide polymorphism at genome position 6446 in clinical samples from patient no. 4

	Day after onset of symptoms						
	5	6	7	8	9	10	11
Swab	A		A				
Sputum		G	G	G	G>A		
Stool			G>A	A=G	A=G	G>A	A

a detection rate of 39.93%. The last swab sample that tested positive was taken on day 28 after the onset of symptoms. The average viral load in sputum was 7.00×10^6 copies per ml, with a maximum of 2.35×10^9 copies per ml.

Because swab samples had limited sensitivity for the initial diagnosis of cases of SARS[13,14], we analysed the first paired swab and sputum samples taken on the same occasion from seven patients. All samples were taken between 2 and 4 days after the onset of symptoms. In two cases, swab samples had virus concentrations that were clearly higher than those in sputum samples, as indicated by a difference of >3 in the threshold cycle (C_t) value. The opposite was true in two other cases, and the remaining three cases had similar concentrations in both sample types.

None of 27 urine samples and none of 31 serum samples tested positive for RNA from SARS-CoV2.

To understand infectivity, we attempted live virus isolation on multiple occasions from clinical samples (Fig. 1d). Whereas the virus was readily isolated during the first week of symptoms from a considerable fraction of samples (16.66% of swabs and 83.33% of sputum samples), no isolates were obtained from samples taken after day 8 in spite of ongoing high viral loads.

Virus isolation from stool samples was never successful, irrespective of viral RNA concentration, on the basis of a total of 13 samples taken between day 6 and day 12 from 4 patients. The success of virus isolation also depended on viral load: samples that contained $<10^6$ copies per ml (or copies per sample) never yielded an isolate. For swab and sputum samples, interpolation based on a probit model was done to obtain laboratory-based infectivity criteria for the discharge of patients (Fig. 1e, f).

High viral loads and successful isolation from early throat swabs suggested potential virus replication in tissues of the upper respiratory tract. To obtain proof of active virus replication in the absence of histopathology, we conducted RT–PCR tests to identify viral subgenomic mRNAs directly in clinical samples (Extended Data Fig. 1). Viral subgenomic mRNA is transcribed only in infected cells and is not packaged into virions, and therefore indicates the presence of actively infected cells in samples. Levels of viral subgenomic mRNA were compared against viral genomic RNA in the same sample. In sputum samples taken on day 4 to day 9, during which time active replication in sputum was obvious in all patients as per longitudinal viral load courses (as described in 'Viral load, antibody response and clinical course'), the ratios of mean normalized subgenomic mRNA per genome were about 0.4% (Fig. 1g). A decline occurred from day 10 to day 11. In throat swabs, all samples taken up to day 5 were in the same range, whereas no subgenomic mRNA was detectable in swabs thereafter. Together, these data indicate the active replication of SARS-CoV-2 in the throat during the first five days after the onset of symptoms. No (or only minimal) indications of replication in stool were obtained by the same method (Fig. 1g).

During our study, we sequenced full virus genomes from all patients. A G6446A exchange was first detected in one patient, and later transmitted to other patients in the cluster[11]. In the first patient, this mutation was found in a throat swab while a sputum sample from the same day showed the original allele (G6446). The single-nucleotide

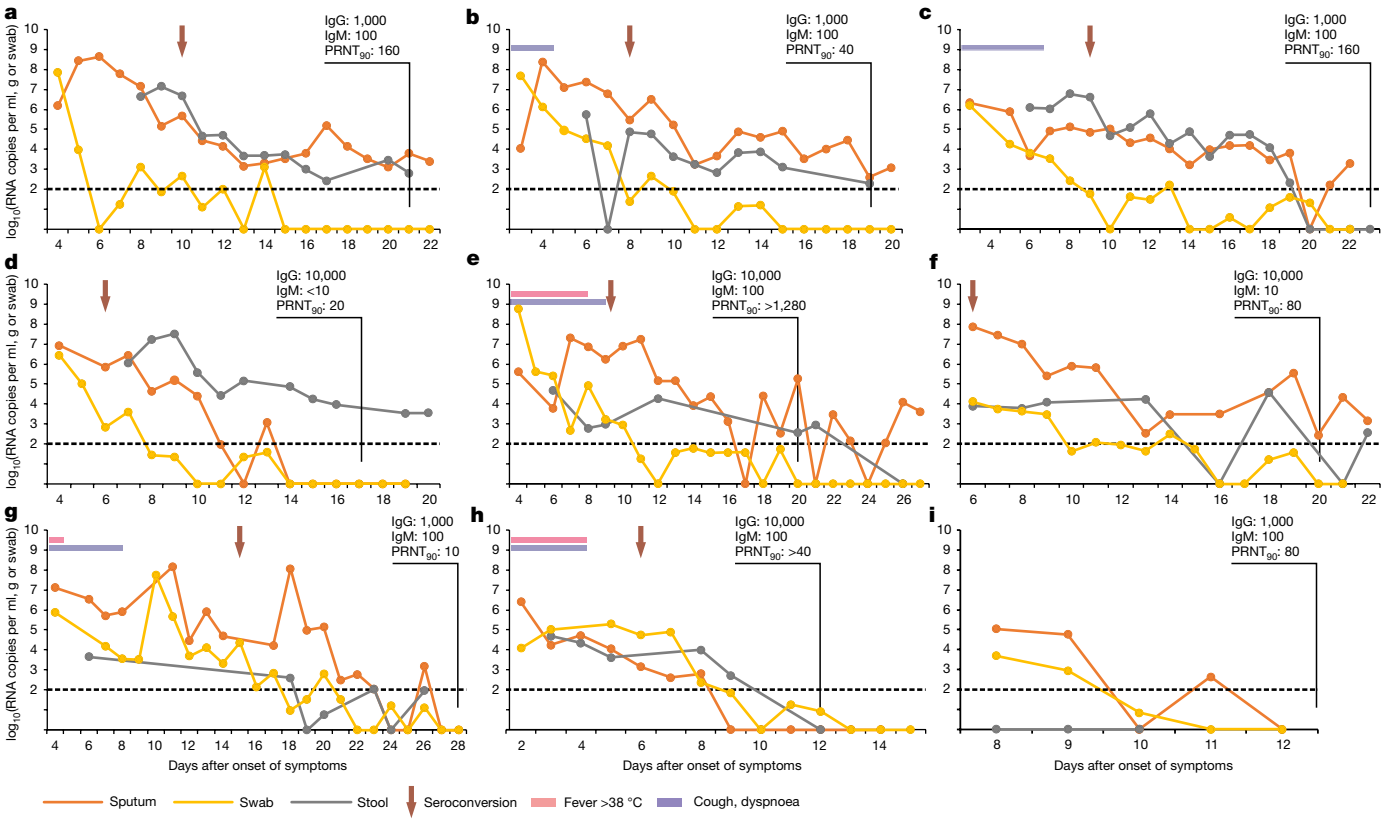


Fig. 2 | Viral load kinetics, seroconversion and clinical observations in individual cases. a–i, The panels correspond to patients no. 1 (**a**), 2 (**b**), 3 (**c**), 4 (**d**), 7 (**e**), 8 (**f**), 10 (**g**), 14 (**h**) and 16 (**i**) in a previous publication[11]. Dotted lines, limit of quantification. Experiments were performed in duplicate and the data presented are the mean of results obtained by two laboratories independently. $PRNT_{90}$, serum dilution that causes viral plaque reduction of 90%.

polymorphism was analysed by RT–PCR and Sanger sequencing in all sequential samples available from that patient (Table 1). The presence of separate genotypes in throat swabs and sputum strongly supported our suspicion of independent virus replication in the throat, rather than passive shedding to the throat from the lung.

Viral load, antibody response and clinical course

Daily measurements of viral load in sputum, pharyngeal swabs and stool are summarized in Fig. 2. In general, the concentrations of viral RNA were very high in initial samples. In all patients except one, the concentration of viral RNA in throat swabs seemed to be already on the decline at the time of first presentation. Viral RNA concentrations in sputum declined more slowly, with a peak during the first week of symptoms in three out of eight patients. Viral RNA concentrations in stools were also high. In many cases, the course of viral RNA concentration in stools seemed to reflect the course in sputum (Fig. 2a–c). In only one case did independent replication in the intestinal tract seem obvious from the course of stool RNA excretion (Fig. 2d). Whereas symptoms mostly waned until the end of the first week (Table 2), viral RNA remained detectable in throat swabs well into the second week. Stool and sputum samples remained RNA-positive over three weeks in six of the nine patients, in spite of full resolution of symptoms.

All cases had comparatively mild courses (Table 2). The two patients who showed some signs of lung infection were the only cases in which sputum viral loads showed a late and high peak around day 10 or 11, whereas sputum viral loads were on the decline by this time in all other patients (Fig. 2f, g). Of note, four out of nine patients showed a loss of taste and olfactory sensation, and described this loss to be stronger and more long-lasting than in common cold diseases.

Seroconversion was detected by IgG and IgM immunofluorescence using cells that express the spike protein of SARS-CoV-2 and a virus neutralization assay using SARS-CoV-2 (Table 3, Extended Data Fig. 2). Seroconversion in 50% of patients occurred by day 7, and in all patients by day 14 (Fig. 1d). No viruses were isolated after day 7. All patients showed detectable neutralizing antibodies, the titres of which did not suggest close correlation with clinical courses. Of note, patient no. 4, who showed the lowest virus neutralization titre at end of week 2, seemed to shed virus from stool over a prolonged time (Fig. 2d). Results from the differential recombinant immunofluorescence assay indicated cross-reactivity or cross-stimulation against the four endemic human coronaviruses in several patients (Extended Data Table 1).

Conclusions

The clinical courses in the patients under study—all of whom were young- to middle-aged professionals without notable underlying disease—were mild. Apart from one patient, all cases were first tested when symptoms were still mild or in the prodromal stage (a period in which most patients would present once there is general awareness of a circulating pandemic disease[5]). Diagnostic testing suggests that simple throat swabs will provide sufficient sensitivity at this stage of infection. This is in stark contrast to SARS; for instance, only 38 of 98 nasal or nasopharyngeal swab samples tested positive by RT–PCR in patients with SARS in Hong Kong[15]. Viral load also differs considerably between SARS and COVID-19. For SARS, it took 7 to 10 days after the onset of symptoms until peak RNA concentrations (of up to 5×10^5 copies per swab) were reached[13,14]. In the present study, peak concentrations were reached before day 5, and were more than 1,000 times higher. Successful isolation of live virus from throat swabs is another notable difference between COVID-19 and SARS, for which such isolation was

Table 2 | Clinical characteristics of all patients

Patient ID no.	Comorbidity	Initial symptoms	Later symptoms	ANC per µl	ALC per µl	CRP (mg l^{-1})	LDH (U l^{-1})
1	Hypothyroidism	Cough, fever, diarrhoea	Diarrhoea	4,870	1,900	46	197
2	None	Sinusitis, cephalgia, cough	Hyposmia, ageusia	3,040	1,200	4.9	182
3	COPD	Arthralgia, sinusitis, cough	Dysosmia, dysgeusia	5,040	2,600	1.3	191
4	None	Otitis, rhinitis	Hyposmia, hypogeusia	2,420	2,220	5.9	149
7	Hypercholesterolaemia	Rhinitis, cough	Fever, dyspnoea, hyposmia, hypogeusia	4,690	900	4.9	209
8	None	Sinusitis, cough		2,500	1,600	1.7	203
10	None	Sinusitis, cough	Fever, cough	2,350	700	7.8	220
14	None	Fever, cough, diarrhoea		5,040	1,500	9.8	220
16	None	None		4,620	900	0.5	201

ALC, absolute lymphocyte count; ANC, absolute neutrophil count; CRC, C-reactive protein; COPD, chronic obstructive pulmonary disease; LDH, lactate dehydrogenase.

rarely successful[16–18]. This suggests active virus replication in tissues of the upper respiratory tract, where SARS-CoV is not thought to replicate in spite of detectable ACE2 expression[19,20]. At the same time, the concurrent use of ACE2 as a receptor by SARS-CoV and SARS-CoV-2 corresponds to a highly similar excretion kinetic in sputum, with active replication in the lung. SARS-CoV was previously found[13] in sputum at mean concentrations of $1.2–2.8 \times 10^6$ copies per ml, which corresponds to observations made here.

Whereas proof of replication by histopathology is awaited, extended tissue tropism of SARS-CoV-2 with replication in the throat is strongly supported by our studies of cells that transcribe subgenomic mRNA in throat swab samples, particularly during the first 5 days of symptoms. Notable additional evidence for independent replication in the throat is provided by sequence findings in one patient, who consistently showed a distinct virus in the throat as opposed to the lung. In addition, the disturbance of gustatory and olfactory senses points at an infection of the tissues of the upper respiratory tract.

Critically, the majority of patients in the present study seemed to be beyond their shedding peak in samples from the upper respiratory tract when they were first tested, whereas the shedding of infectious virus in sputum continued throughout the first week of symptoms. Together, these findings suggest a more efficient transmission of SARS-CoV-2 than SARS-CoV, through active pharyngeal viral shedding at a time at which symptoms are still mild and typical of infections of the upper respiratory tract. Later in the disease, COVID-19 resembles SARS in terms of replication in the lower respiratory tract. Of note, the two patients who showed some symptoms of the lungs being affected showed a prolonged viral load in sputum. Our study is limited, in that no severe cases were observed. Future studies that include severe cases should look at the prognostic value of an increase of viral load beyond the end of week 1, potentially indicating an aggravation of symptoms.

One of the most interesting hypotheses to explain the potential extension of tropism to the throat is the presence of a polybasic furin-type cleavage site at the S1–S2 junction in the SARS-CoV-2 spike protein that is not present in SARS-CoV[17]. The insertion of a polybasic cleavage site in the S1–S2 region in SARS-CoV has previously been shown to lead to a moderate, but discernible, gain-of-fusion activity that might result in increased viral entry in tissues with a low density of ACE2 expression[21].

The combination of very high concentrations of virus RNA and the occasional detection of cells in stools that contain subgenomic mRNA indicate active replication in the gastrointestinal tract. Active replication is also suggested by a much higher detection rate compared to the Middle East respiratory system coronavirus (MERS-CoV), for which stool-associated RNA was found in only 14.6% of samples from 37 patients hospitalized in Riyadh (Saudi Arabia)[22,23]. If SARS-CoV-2 was only passively present in the stool (such as after swallowing respiratory secretions), similar detection rates as for MERS-CoV would be expected. Replication in the gastrointestinal tract is also supported by analogy with SARS-CoV, which was regularly excreted in stool (from which it could be isolated in cell culture[24]). Our failure to isolate live SARS-CoV-2 from stools may be due to the mild courses of cases, with only one case showing intermittent diarrhoea. In China, diarrhoea was seen in only 2 of 99 cases[25]. Further studies should therefore address whether SARS-CoV-2 shed in stools is rendered noninfectious though contact with the gut environment. Our initial results suggest that measures to contain viral spread should aim at droplet-, rather than fomite-, based transmission.

The prolonged viral shedding in sputum is relevant not only for the control of infections in hospitals, but also for discharge management. In a situation characterized by a limited capacity of hospital beds in infectious disease wards, there is pressure for early discharge after treatment. On the basis of the present findings, early discharge with ensuing home isolation could be chosen for patients who are beyond day 10 of symptoms and have less than 100,000 viral RNA copies per ml of sputum. Both criteria predict that there is little residual risk of infectivity, on the basis of cell culture.

The serological courses of all patients suggest a timing of seroconversion similar to, or slightly earlier than, in SARS-CoV infection[18]. Seroconversion in most cases of SARS occurred during the second week of symptoms. As in SARS and MERS, IgM was not detected considerably earlier than IgG in immunofluorescence; this might in part be due to technical reasons, as the higher avidity of IgG antibodies outcompetes IgM for viral epitopes in the assay. IgG depletion can only partially alleviate this effect. Because immunofluorescence assay is a

Table 3 | IgG and IgM immunofluorescence titres against SARS-CoV-2, from all patients

Patient ID no.	Initial serum		Final serum				
	Day after onset	IgG	Day after onset	IgG	IgM	$PRNT_{90}$	$PRNT_{50}$
1	5	<10	21	1,000	100	160	>640
2	4	<10	19	1,000	100	40	320
3	3	<10	23	1,000	100	160	>640
4	5	<10	17	10,000	<10	20	160
7	6	<10	20	10,000	100	>1,280	>1,280
8	6	10	20	10,000	10	80	>320
10	6	<10	28	1,000	10	10	>40
14	NA	NA	12	10,000	100	>40	>40
16	NA	NA	13	1,000	100	80	>320

NA, not applicable; $PRNT_{50}$, serum dilution that causes viral plaque reduction of 50%.

labour-intensive method, enzyme-linked immunosorbent assay tests should be developed as a screening test. Neutralization testing is necessary to rule out cross-reactive antibodies directed against endemic human coronaviruses. On the basis of the frequently low neutralizing antibody titres observed in coronavirus infection[26,27], we have here developed a particularly sensitive plaque-reduction neutralization assay. Considering the titres we observed, a simpler microneutralization test format is likely to provide sufficient sensitivity in routine application and population studies.

When aligned to viral load courses, it seems there is no abrupt virus elimination at the time of seroconversion. Rather, seroconversion early in week 2 coincides with a slow but steady decline of viral load in sputum. Whether properties such as the glycosylation pattern at critical sites of the glycoprotein have a role in the attenuation of the neutralizing antibody response needs further clarification. In any case, vaccine approaches targeting mainly the induction of antibody responses should aim to induce particularly strong antibody responses to be effective.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2196-x.

1. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med*. **382**, 727–733 (2020).
2. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol*. **5**, 536–544 (2020).
3. WHO. *Report of the WHO–China Joint Mission on Coronavirus Disease 2019 (COVID-19)* https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf (WHO, 2020).
4. Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280 (2020).
5. Leung, G. M. et al. The epidemiology of severe acute respiratory syndrome in the 2003 Hong Kong epidemic: an analysis of all 1755 patients. *Ann. Intern. Med*. **141**, 662–673 (2004).
6. Rothe, C. et al. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *N. Engl. J. Med*. **382**, 970–971 (2020).
7. Holshue, M. L. et al. First case of 2019 novel coronavirus in the United States. *N. Engl. J. Med*. **382**, 929–936 (2020).
8. Hoehl, S. et al. Evidence of SARS-CoV-2 infection in returning travelers from Wuhan, China. *N. Engl. J. Med*. **382**, 1278–1280 (2020).
9. Zou, L. et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N. Engl. J. Med*. **382**, 1177–1179 (2020).
10. Young, B. E. et al. Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore. *J. Am. Med. Assoc*. **323**, 1488–1494 (2020).
11. Böhmer, M. et al. Outbreak of COVID-19 in Germany resulting from a single travel-associated primary case. *Lancet Infect. Dis.* https://doi.org/10.1016/S1473-3099(20)30314-5 (2020).
12. Corman, V. M. et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT–PCR. *Euro Surveill*. **25**, 1–8 (2020).
13. Drosten, C. et al. Evaluation of advanced reverse transcription–PCR assays and an alternative PCR target region for detection of severe acute respiratory syndrome-associated coronavirus. *J. Clin. Microbiol*. **42**, 2043–2047 (2004).
14. Peiris, J. S. et al. Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia: a prospective study. *Lancet* **361**, 1767–1772 (2003).
15. Poon, L. L. et al. Detection of SARS coronavirus in patients with severe acute respiratory syndrome by conventional and real-time quantitative reverse transcription-PCR assays. *Clin. Chem*. **50**, 67–72 (2004).
16. Ksiazek, T. G. et al. A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med*. **348**, 1953–1966 (2003).
17. Drosten, C. et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med*. **348**, 1967–1976 (2003).
18. Peiris, J. S. et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319–1325 (2003).
19. Bertram, S. et al. Influenza and SARS-coronavirus activating proteases TMPRSS2 and HAT are expressed at multiple sites in human respiratory and gastrointestinal tracts. *PLoS ONE* **7**, e35876 (2012).
20. Xu, H. et al. High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa. *Int. J. Oral Sci*. **12**, 8 (2020).
21. Belouzard, S., Chu, V. C. & Whittaker, G. R. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proc. Natl Acad. Sci. USA* **106**, 5871–5876 (2009).
22. Corman, V. M. et al. Viral shedding and antibody response in 37 patients with Middle East respiratory syndrome coronavirus infection. *Clin. Infect. Dis*. **62**, 477–483 (2016).
23. Zhou, J. et al. Human intestinal tract serves as an alternative infection route for Middle East respiratory syndrome coronavirus. *Sci. Adv*. **3**, eaao4966 (2017).
24. Leung, W. K. et al. Enteric involvement of severe acute respiratory syndrome-associated coronavirus infection. *Gastroenterology* **125**, 1011–1017 (2003).
25. Chen, N. et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **395**, 507–513 (2020).
26. Drosten, C. et al. Transmission of MERS-coronavirus in household contacts. *N. Engl. J. Med*. **371**, 828–835 (2014).
27. Müller, M. A. et al. Presence of Middle East respiratory syndrome coronavirus antibodies in Saudi Arabia: a nationwide, cross-sectional, serological study. *Lancet Infect. Dis*. **15**, 559–564 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Clinical samples and viral load conversion

Sputum and stool samples were taken and shipped in native conditions. Oro- and nasopharyngeal throat swabs were preserved in 3 ml of viral transport medium. Viral loads in sputum samples were projected to RNA copies per ml, in stool samples to copies per g and in throat swabs to copies per 3 ml, assuming that all sample components were suspended in 3 ml viral transport medium. For swab samples suspended in less than 3 ml viral transport medium, this conversion was adapted to represent copies per whole swab. An aggregated overview of samples received per day after the onset of disease from all patients is shown in Fig. 1a.

RT–PCR for SARS-CoV-2 and other respiratory viruses

RT–PCR used targets in the *E* and *RdRp* genes as previously described[12]. Both laboratories used a pre-formulated oligonucleotide mixture (Tib-Molbiol) to make the laboratory procedures more reproducible. All patients were also tested for other respiratory viruses, including HCoV-HKU1, HCoV-OC43, HCoV-NL63 and HCoV-229E, influenza virus A, influenza virus B, rhinovirus, enterovirus, respiratory syncytial virus, human parainfluenza viruses 1–4, human metapneumovirus, adenovirus and human bocavirus using LightMix-Modular Assays (Roche). Additional technical details are provided in Supplementary Methods section 1.

Virus isolation

Virus isolation was done in two laboratories on Vero E6 cells. In brief, 100 µl of suspended, cleared and filtered clinical sample was mixed with an equal volume of cell culture medium. Supernatant was collected after 0, 1, 3 and 5 days and used in RT–PCR analysis. Additional technical details are provided in Supplementary Methods section 2a.

Serology

We performed recombinant immunofluorescence assays to determine the specific reactivity against recombinant spike proteins in VeroB4 cells, as previously described[26,28]. This assay used a cloned coronavirus spike protein from HCoV-229E, HCoV-NL63, HCoV-OC43, HCoV-HKU1 or SARS-CoV-2. The screening dilution was 1:10. Plaque reduction neutralization tests were done essentially as previously described for MERS-CoV[26]. Serum dilutions causing plaque reductions of 90% ($PRNT_{90}$) and 50% ($PRNT_{50}$) were recorded as titres. Additional technical details are provided in Supplementary Methods section 2b, c.

Statistical analyses

Statistical analyses were done using SPSS software (version 25) or GrapPad Prism (version 8).

Ethical approval statement

All patients provided informed consent for the use of their data and clinical samples for the purposes of the present study. Institutional review board clearance for the scientific use of patient data has been granted to the treating institution by the ethics committee at the Medical Faculty of the Ludwig Maximillians Universität Munich (vote 20-225 KB).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sequence data are available in Gisaid under accession number EPI_ISL_406862. All other data are available from C.D. upon reasonable request.

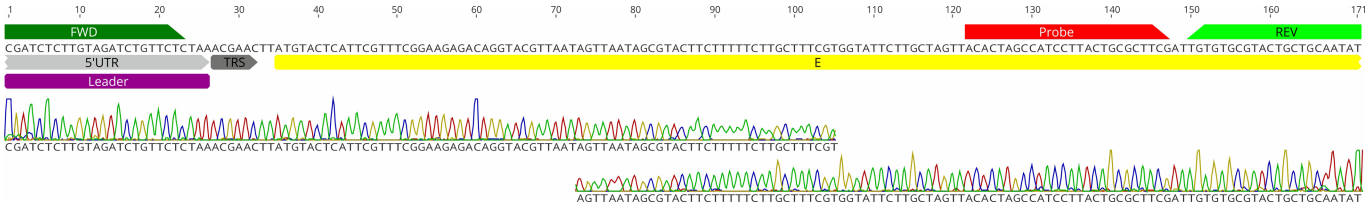
28. Corman, V. M. et al. Assays for laboratory confirmation of novel human coronavirus (hCoV-EMC) infections. *Euro Surveill.* **17**, 20334 (2012).

Acknowledgements This work was funded by grants from the German Ministry of Research (01KI1723A) and the European Union (602525) to C.D. as well as by the German Bundeswehr Medical Service Biodefense Research Program. The funders had no role in study design, data collection and analysis or decision to publish. We thank P. Mackeldanz, E. Möncke-Buchner, A. Richter, M. Schmidt and J. Beheim-Schwarzbach for technical assistance.

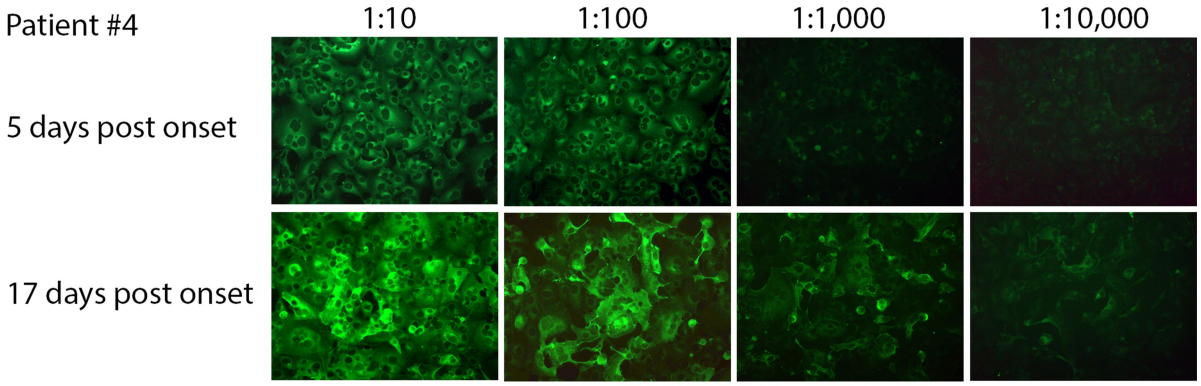
Author contributions R.W. and V.M.C. planned and supervised laboratory testing, and evaluated data. W.G. and M.S. managed patients and evaluated clinical data. S.Z., T.B., S.B., J.S., R.E. and K.Z. performed laboratory testing. M.A.M. managed serological laboratory testing. D.N. managed and performed virus isolation studies. T.C.J. analysed sequences and population-specific polymorphisms. P.V. managed laboratory testing. C.R. managed initial patient contacts. M.H. managed initial patient contacts and evaluated clinical data. C.D. designed and supervised laboratory studies, and wrote the manuscript. C.W. designed and supervised clinical management and clinical data.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2196-x.
Correspondence and requests for materials should be addressed to C.D. or C.W.
Peer review information *Nature* thanks Peter Openshaw and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Sequence analysis of *E* gene subgenomic mRNA. The leader sequence (purple), putative transcription regulatory sequences (TRS) (grey) and nucleotides coding for the 5′-proximal part of the *E* gene (yellow box) are shown. PCR primer binding sites used for amplification and RT–PCR detection are shown as green arrows, and the 5′-nuclease PCR probe is shown as a red arrow.



Extended Data Fig. 2 | Recombinant SARS-CoV-2-spike-based immunofluorescence test shows seroconversion of patient no. 4. Representative outcome of a recombinant immunofluorescence test using serum dilutions 1:10, 1:100, 1:1,000 and 1:10,000 of patient no. 4 at 5 and 17 days after the onset of symptoms. Secondary detection was done by using a goat-anti human immunoglobulin labelled with Alexa Fluor 488 (shown in green). The experiment was performed in duplicate.

Extended Data Table 1 | IgG immunofluorescence titres against endemic human coronaviruses

Patient ID	*Primary serum*					*Final serum*				
	Day p.o.	OC43	NL63	HKU1	229E	Day p.o.	OC43	NL63	HKU1	229E
#1	5	1,000	1,000	1,000	100	15	1,000	1,000	1,000	100
#2	4	1,000	1,000	100	100	13	**10,000**	100	**1,000**	10
#3	3	10,000	100	1,000	1000	16	10,000	**1,000**	**10,000**	1,000
#4	5	1,000	100	100	100	17	**10,000**	10	**1,000**	100
#7	6	1,000	100	1,000	1000	13	**10,000**	**1,000**	**10,000**	**10,000**
#8	6	1,000	100	1,000	1000	10	**10,000**	**1,000**	**10,000**	100
#10	6	1,000	100	100	1000	11	**10,000**	**1,000**	100	1,000
#14	na	na	na	na	na	5	100	100	100	100
#16	na	na	na	na	na	13	10,000	1,000	1,000	100

p.o., post onset; na, not available. Increases of titre through the final serum are indicated by reciprocal titres in bold.

nature research

Corresponding author(s): *Double-blind peer review submissions: write DBPR and your manuscript number here instead of author names.*

Last updated by author(s): *YYYY-MM-DD*

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Microsoft Excel,
Data analysis	SPSS, GraphPad Prism, Geneious

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data upon which figures and conclusions are based can be obtained from the corresponding author upon reasonable request. These data cannot be published in an open source because their interpretation may affect aspects of patient privacy.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Nine Patients.
Data exclusions	None.
Replication	Testing by two different laboratories.
Randomization	None.
Blinding	None.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Secondary anti-human IgG reagents (Euroimmun).
Validation	Immunofluorescence.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	In-house collection with reference to ATCC or DSZM.
Authentication	Functional testing for IFN locus inactivation.
Mycoplasma contamination	Regular testing.
Commonly misidentified lines (See ICLAC register)	N/A

Human research participants

Policy information about studies involving human research participants

Population characteristics	N.A.
Recruitment	Clinical admission due to symptoms, contact history, and positive initial test.
Ethics oversight	Research ethics board of Ludwig Maximillians University Munich; informed consent to scientific use and publication of anonymized data by each patient.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	N.A.
Study protocol	No study protocol emergency admissions of patients with new disease.
Data collection	Data collection at treating hospital, and two laboratories as identifed in affiliations list.
Outcomes	Laboratory and clinical status. No explicit outcome measure.

Article

The stepwise assembly of the neonatal virome is modulated by breastfeeding

https://doi.org/10.1038/s41586-020-2192-1

Received: 12 July 2019

Accepted: 14 February 2020

Published online: 15 April 2020

Check for updates

Guanxiang Liang[1,2], Chunyu Zhao[2], Huanjia Zhang[2], Lisa Mattei[2], Scott Sherrill-Mix[1], Kyle Bittinger[2], Lyanna R. Kessler[1], Gary D. Wu[3], Robert N. Baldassano[2], Patricia DeRusso[2], Eileen Ford[2], Michal A. Elovitz[4], Matthew S. Kelly[5], Mohamed Z. Patel[6], Tiny Mazhani[6], Jeffrey S. Gerber[7], Andrea Kelly[8], Babette S. Zemel[2] & Frederic D. Bushman[1]✉

The gut of healthy human neonates is usually devoid of viruses at birth, but quickly becomes colonized, which—in some cases—leads to gastrointestinal disorders[1–4]. Here we show that the assembly of the viral community in neonates takes place in distinct steps. Fluorescent staining of virus-like particles purified from infant meconium or early stool samples shows few or no particles, but by one month of life particle numbers increase to 10^9 per gram, and these numbers seem to persist throughout life[5–7]. We investigated the origin of these viral populations using shotgun metagenomic sequencing of virus-enriched preparations and whole microbial communities, followed by targeted microbiological analyses. Results indicate that, early after birth, pioneer bacteria colonize the infant gut and by one month prophages induced from these bacteria provide the predominant population of virus-like particles. By four months of life, identifiable viruses that replicate in human cells become more prominent. Multiple human viruses were more abundant in stool samples from babies who were exclusively fed on formula milk compared with those fed partially or fully on breast milk, paralleling reports that breast milk can be protective against viral infections[8–10]. Bacteriophage populations also differed depending on whether or not the infant was breastfed. We show that the colonization of the infant gut is stepwise, first mainly by temperate bacteriophages induced from pioneer bacteria, and later by viruses that replicate in human cells; this second phase is modulated by breastfeeding.

To investigate early-life viral colonization, we first analysed stool samples from 20 healthy infants longitudinally (Supplementary Table 1). Samples included meconium and/or early stool samples collected 0–4 days after birth (median of 17 h after birth, range 11–152 h; hereafter 'month 0') and stool samples collected at 1 and 4 months of life. The cohort consisted of self-identified African-American mothers from an urban US setting and their infants. As an initial step, virus-like particles (VLPs) were purified from meconium or stool and stained with SYBR gold, which binds nucleic acids. VLPs were subsequently visualized by epifluorescence microscopy (Fig. 1a). VLPs were undetectable in most of the meconium samples; only 3 out of 20 samples had detectable counts of VLPs (Fig. 1b). By month 1, most samples were positive, and VLP counts averaged 1.6×10^9 per gram of stool; values at month 4 were similar to the 1-month samples. We also tested the VLP counts from 12 2–5-year-old children, which had an average of 9.4×10^8 per gram of stool; these results were not distinguishable from month-1 and month-4 samples ($P = 0.48$, Wilcoxon rank-sum test). This number is also close to that reported for adults[5–7]. We therefore conclude that the high VLP counts seen in 1-month-old infants typically persist into adulthood.

To characterize bacterial content, DNA was purified from whole meconium or stool samples and analysed by qPCR to quantify the copy numbers of the bacterial 16S rRNA gene (Fig. 1c). Some published studies have suggested that microbial colonization of the infant begins in utero[11], but recent studies indicate that colonization more likely begins with the rupture of membranes and delivery[12–14]. Quantification showed low or undetectable levels of bacterial 16S rRNA genes for 14 of the 20 meconium/early stool samples from month 0, and relatively low levels for the other 6 (median, 3.3×10^6 per gram of stool). By contrast, for months 1 and 4, most samples were positive for 16S rRNA gene sequences, with a median value of 3.1×10^8 per gram of stool). To analyse the early-life microbiome further, total DNA from each sample was subjected to metagenomic sequencing. For the month-0 samples, many were dominated by human DNA, which is characteristic of the neonatal gut before colonization (Extended Data Fig. 1a). Some of the month-0 samples also contained bacterial DNA, indicating early colonization and/or reagent contamination. Levels of human DNA decreased with time after delivery ($P = 0.04$, Spearman's rank-order correlation $\rho = -0.45$; Extended Data Fig. 1b), consistent with bacterial

[1]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]Division of Gastroenterology, Hepatology, and Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [3]Division of Gastroenterology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [4]Maternal and Child Health Research Center, Department of Obstetrics and Gynecology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [5]Division of Pediatric Infectious Diseases, Duke University, Durham, NC, USA. [6]Department of Paediatric and Adolescent Health, Faculty of Medicine, University of Botswana, Gaborone, Botswana. [7]Division of Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [8]Division of Endocrinology and Diabetes, Children's Hospital of Philadelphia, Philadelphia, PA, USA. ✉e-mail: bushman@pennmedicine.upenn.edu

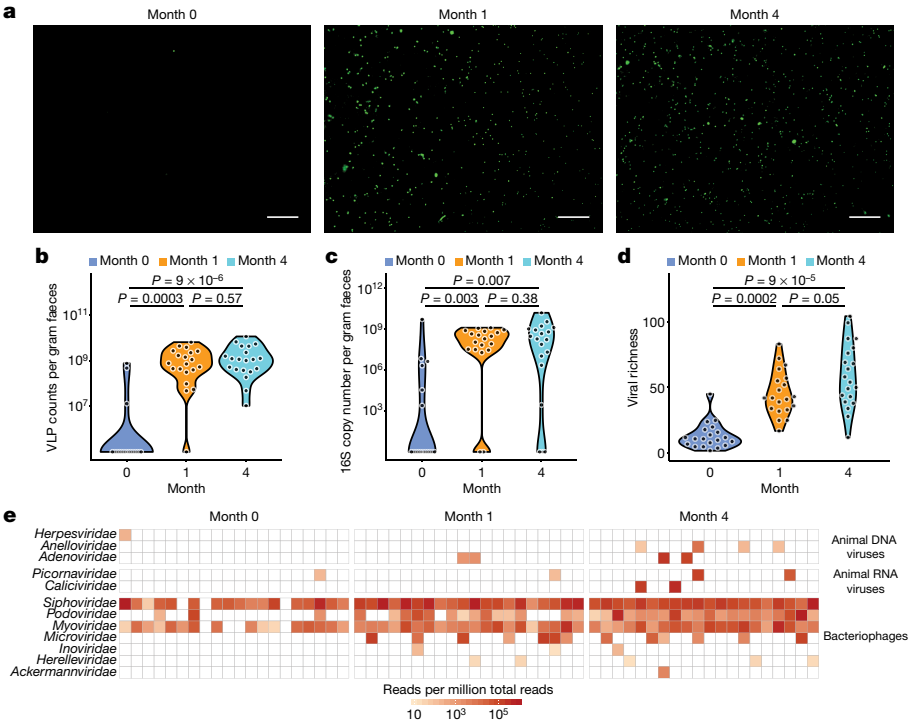


Fig. 1 | Detection and characterization of VLPs in infant gut samples. **a**, Representative fields of fluorescently stained VLPs from infant stool sampled at months 0, 1 and 4. Scale bars, 10 μm. **b**, Quantification of VLP counts per gram faeces. The minimum level of quantification was 6.6×10^6 particles per gram. Per sample, 5–10 fields were quantified. **c**, Copy numbers of bacterial 16S rRNA genes analysed using qPCR. The minimum level of quantification was 2,000 copies per gram. **d**, VLP richness was assessed using VLP metagenomic sequence data. Sequence reads were assembled into contigs, and contigs with viral characteristics (at least 50% of open-reading frames annotated as viral) were enumerated. Viral species were called present if at least 10 reads per million reads from one sample aligned to that contig. **e**, Taxonomic assignments of VLP sequences. Reads were associated with viral lineages based on the annotation of viral contigs. In **b**–**d**, violin plots represent the distribution of the individual datasets; samples were compared using two-sided Wilcoxon signed-rank tests.

colonization. Bacterial DNA predominated in month-1 and month-4 samples (Extended Data Fig. 1a). The bacterial richness and diversity at month 0 was lower compared to month-1 and month-4 samples. Early bacterial colonizers included Proteobacteria, Actinobacteria, Bacteroidetes and Firmicutes (Extended Data Fig. 1c–e), consistent with previous studies[2,15].

To investigate the origin of the early life virome, DNA and RNA were purified from preparations of VLPs from stool from each of the three time points and characterized by metagenomic sequencing. After filtering out human DNA, we assembled sequence reads into contigs and annotated open-reading frames. Previous literature indicates that many gut viruses are uncharacterized bacteriophages (hereafter phages)[1,16–19], which are challenging to identify in metagenomic sequence data because the proportion of phage genomes in databases is small compared to the number of global phage types. Viruses that infect human cells are more fully characterized and thus more readily recognizable. To address this challenge, we required that half of all reading frames within a contig were annotated as viral to assign that contig as a virus. Quantification of viral species richness showed low values at month 0, but higher richness after 1 and 4 months (Fig. 1d). After taxonomic assignment of viral contigs, we found that, despite the difficulty of identifying phages, the majority of VLP classifications were in fact phage families (Fig. 1e). Most were from DNA phages, consistent with the reported rarity of RNA phages[20]. For DNA phages, an average of 31% of reads could be assigned as viral at month 1 and 38% at month 4. The nature of the remainder is unknown but some reads probably represent unstudied phages. Values were lower at month 0 (11%), which probably reflects a relative increase in background due to the low numbers of recovered particles (Extended Data Fig. 2a–h).

To assess community interactions, we compared bacterial abundances from 16S rRNA gene qPCR data, bacterial richness and bacterial diversity from sequence data with VLP counts, viral richness and viral diversity, and found strong positive correlations (Extended Data Fig. 3a, b and Supplementary Table 2).

For the minority of viruses detected that are known to replicate in human cells (Fig. 1e), at month 0 a single sample was positive for *Herpesviridae* and another for *Picornaviridae*. By month 4, human-cell viruses were more prominent, including *Adenoviridae*, *Anelloviridae*, *Caliciviridae* and *Picornaviridae*.

Either of two modes of phage production could generate the observed VLP populations[21,22]. Lytic phages only grow by infection, replication and lysis (Extended Data Fig. 4a). Previous reports, which focused on older infants and adults, have suggested that lytic growth and predator–prey interactions between phages and bacteria were prominent in early-life communities[2]. Temperate phages can also replicate using a second strategy, which involves the integration of the phage DNA into the host bacterial DNA, followed by quiescent growth as an integrated prophage (Extended Data Fig. 4a). Exposure to an inducing signal causes integrated prophages to excise and resume lytic growth. Induction can also take place at low levels spontaneously[23]. The prophage state is commonly maintained by repressor proteins, which also serve to exclude infection by similar or identical (homoimmune) phage strains[21,22].

To test whether the early-life virome is composed of strictly lytic phages, we purified 24 bacterial strains from the infant gut, including three *Escherichia coli*, three *Klebsiella* and ten *Enterococcus* strains (Supplementary Table 3), and plated virome fractions from the cognate infant back on these bacteria. In no case did virome VLPs form plaques on lawns of these bacteria, thus providing no evidence that

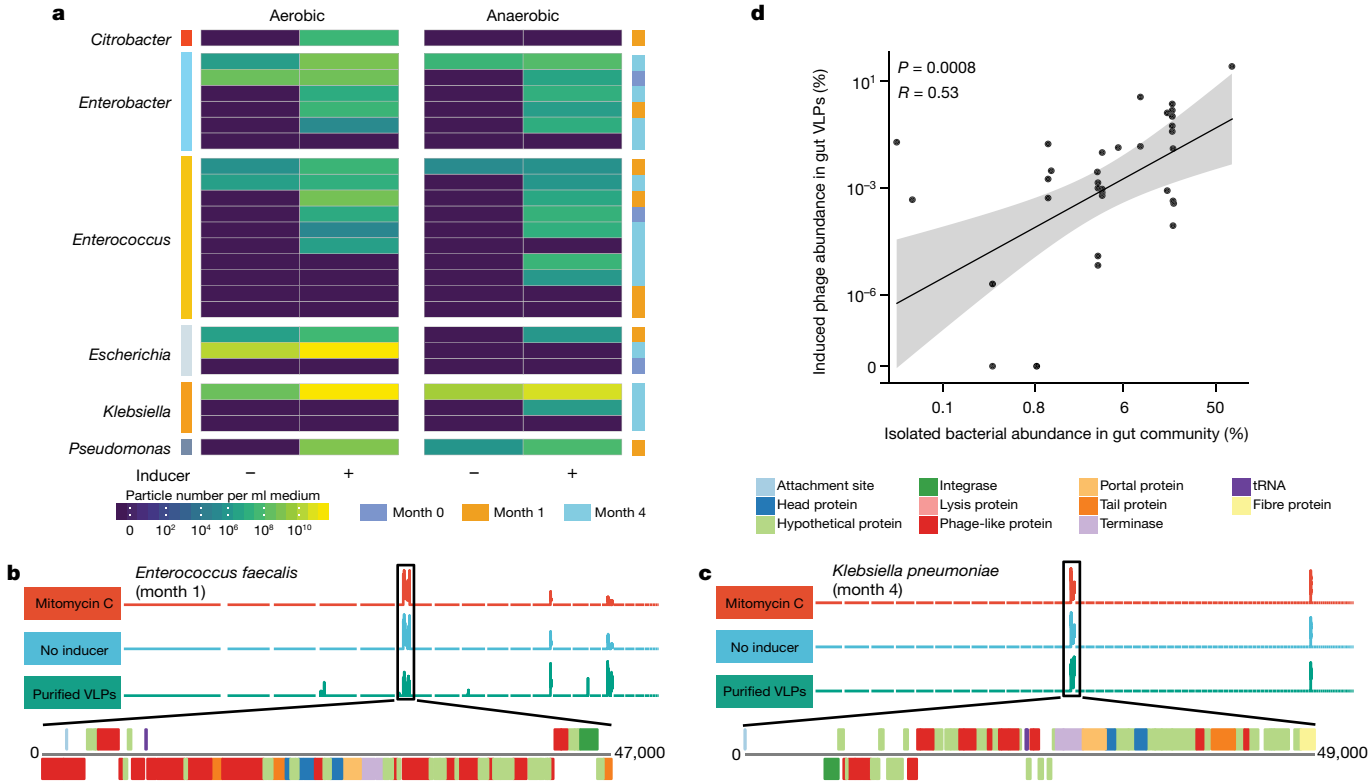


Fig. 2 | Prophage induction as the dominant contributor to the early-life virome. a, Heat map quantifying VLP production from 24 strains isolated from the faeces of the infants studied. The bacterial genera are summarized on the left; columns summarize the numbers of fluorescent particles produced per ml of stationary-phase culture (according to the scale at the bottom). Columns compare particle production with and without an inducer (mitomycin C) as well as growth under aerobic and anaerobic conditions. **b**, Draft genome (horizontal line) of *Enterococcus faecalis* from one of the infants studied, showing the alignment frequency of reads from VLP preparations. Reads were aligned to the bacterial genomes that were generated from VLPs from pure culture after mitomycin treatment (red), from VLPs from pure culture in the absence of any inducer (blue) and from VLPs isolated from the stool of the infant from which the bacterial strain was isolated (green). Peaks indicate the detection of integrated prophages. One putative bacteriophage genome is shown below, with gene types colour-coded as indicated. **c**, As in **b**, but for a *Klebsiella pneumoniae* isolate. **d**, Correlation between the abundance of VLPs present in infant stool and the abundance of the bacteria harbouring those prophages in the same stool sample ($n = 33$ phage contigs from 16 bacterial isolates from month-1 and month-4 strains). The black line shows the linear regression line and the grey-shaded region shows the 95% confidence interval for the slope (two-sided Spearman's rank-order correlation).

lytic replication occurs. Note that temperate phages are not expected to form plaques on host cells that already contain those phages as an integrated prophage due to repressor-mediated homoimmune exclusion[21,22]. Analysis of our DNA VLP contigs using PHACTS, a random forest-based approach to classifying phage lifestyles[24], indicated that most of the genomes more closely resembled temperate phages than lytic phages (Extended Data Fig. 4b).

We next investigated whether virome populations resulted from the induction of prophages by quantifying VLP production from the 24 infant bacterial strains described above. Bacterial strains were analysed for spontaneous VLP production during growth in liquid culture, and for production after induction with the DNA-damaging agent mitomycin C, using the fluorescent staining assay for VLP particles. Experiments were carried out under both aerobic and anaerobic conditions. We detected spontaneous VLP production in 32% of strains. After induction with mitomycin C, 80% of strains produced VLPs (Fig. 2a). In total, 16 strains showed VLP production of at least 10^7 particles per ml under at least one condition. We therefore conclude that bacteria in the infant gut are commonly capable of high-level VLP production following prophage induction.

The hypothesis that prophage induction yields the bulk of VLPs in infant gut samples predicts that VLP sequences found in stool samples should be detectable as integrated prophages in bacterial genome sequences. We therefore sequenced genomes of the 24 infant bacterial strains described above and of VLPs produced from those strains in the presence or absence of mitomycin C (Supplementary Table 3). Prophage sequences could be readily detected in the bacterial genomes, and many of these were detectable both in sequences from the induced VLP samples and in the VLP samples from infant stool. Examples are shown in Fig. 2b, c, in which the VLP sequence reads are shown aligned to bacterial contigs, so that the spikes indicate VLP detection after mitomycin C induction, in the absence of induction and in purified stool VLPs. To test the infant specificity of each community, we mapped the stool VLP reads back to the viral contigs assembled from VLPs induced from the 24 bacterial strains. Although the steps of VLP nucleic acid amplification before sequencing can distort abundances, we nevertheless found that the induced VLPs from each purified bacterial strain were more similar to stool VLP sequences from the infant from which the bacterial strain was isolated than to VLPs from unmatched infants ($P < 0.0001$; Extended Data Fig. 5a), consistent with the production of phages in the infant gut by prophage induction.

In addition, there was a significant positive correlation between the proportion of each bacterium in the infant gut community and the proportion of prophages from that bacterial species in the gut virome of the associated infant ($P = 0.0008$; Spearman's rank-order correlation $\rho = 0.53$; Fig. 2d and Extended Data Fig. 5b). The abundance of the bacterial strains in the gut communities ranged from 0.03% to 99.1%, indicating that in at least some cases a large proportion of the gut community was analysed.

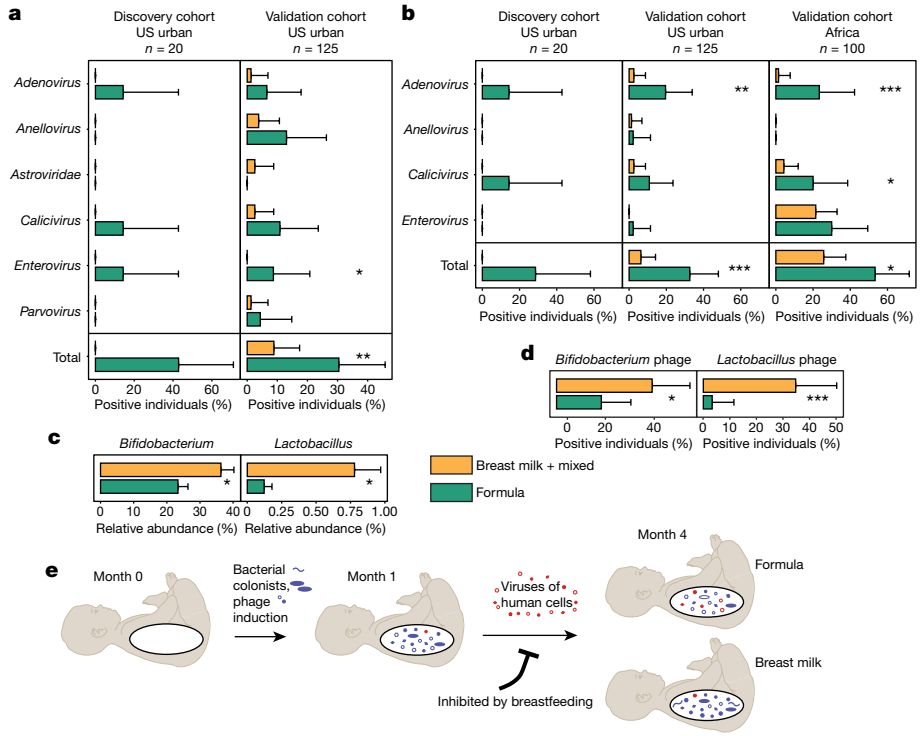


Fig. 3 | Breastfeeding and viral colonization of the infant gut. **a**, Quantification of the percentage of infants who were positive for viruses of human cells in metagenomic virome sequence data. Sample sizes and cohorts studied are indicated at the top. The two feeding types are colour-coded as indicated. The summation over all viral families is shown at the bottom (total). **b**, Comparison of human virus colonization based on feeding type using qPCR. Three technical replicates were compared for each sample. In **a**, **b**, the numbers of infants fed with formula and fed with breast milk or mixed feeding are, respectively, 14 and 6 in the discovery cohort, 46 and 79 in the validation cohort from US urban areas, and 30 and 70 in the validation cohort from Africa. **c**, Abundances of the bacterial genera of *Bifidobacterium* and *Lactobacillus* separated by feeding type. Data were obtained using total microbial shotgun sequencing. **d**, Percentage of infants positive for phages annotated as infecting *Bifidobacterium* or *Lactobacillus*, assessed using VLP metagenomic sequencing. In **c**, **d**, 103 (formula, $n = 59$; breast milk or mixed, $n = 44$) samples from both discovery and validation cohorts were used for which whole-stool shotgun sequencing data were available. **a**, **b**, **d**, Samples were compared using two-sided Fisher's exact tests. Data summarize population proportions and 95% confidence intervals. **c**, Samples were compared using two-sided Wilcoxon rank-sum tests with FDR correction. Data are mean ± s.e.m. **a**–**d**, $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. **e**, Summary of the findings of this study.

The abundant crAssphages, which infect Bacteroides and do not integrate during replication[25,26], were scarce in samples from month 1, but more common by month 4 and in samples from 2–5-year-old children (Extended Data Fig. 6). Evidently this group of lytic phages colonizes children predominantly later in life, potentially reflecting sequential acquisition of Bacteroides strains and later crAssphages from the environment.

Our findings support the idea that prophage induction from pioneer bacteria is the main source of the observed virome community by month 1. This is supported by the findings that: (1) replication of lytic phages was undetectable; (2) many purified bacterial strains from infants produced VLPs at high levels; (3) sequences of genomes from these induced VLPs could be identified as integrated prophages in bacteria isolated from these infants; (4) stool VLP genome sequences could be identified as integrated prophages in the bacterial genomes; (5) VLP abundance in stool was proportional to the abundance of the host bacteria in the same sample; and (6) VLP contigs were annotated primarily as lysogenic phages and not lytic phages.

We then compared features of the VLP data from infant stool samples to metadata on feeding history, mode of delivery, sex and other variables (Supplementary Table 1). We found a strong influence of breastfeeding, which was associated with the lower accumulation of viruses that replicate in human cells. Taking a conservative threshold for detection, requiring coverage of 33% of the viral genome, viruses that infect human cells were only found in those infants who were exclusively fed formula milk (Fig. 3a and Extended Data Fig. 7a). Statistically, this achieved a P value of only 0.11 (Fisher's exact test) owing to the small sample size and unbalanced distribution (Fig. 3a and Extended Data Fig. 7b). Delivery type (spontaneous vaginal delivery versus caesarean section) did achieve statistical significance ($P = 0.01$, Fisher's exact test; Extended Data Fig. 7f, g).

To challenge these conclusions, we analysed a validation cohort of an additional 125 infants, focusing on stool samples taken at 3–4 months of life. These samples were obtained from mixed-race cohorts of urban US infants (Supplementary Table 1). In these samples, delivery mode did not show a significant influence (Extended Data Fig. 7h–j), but a protective effect of breastfeeding was evident—30% of formula-fed babies were positive for viruses that infect human cells, whereas 9% of babies who were fed breast milk or breast milk together with formula were positive ($P = 0.003$, Fisher's exact test; Fig. 3a). Repeating the analysis requiring 0.1% coverage to 60% coverage of viral genomes for scoring detection yielded similarly significant results (Extended Data Fig. 7c, d). A comparison after normalizing for sequencing depth also yielded a significant difference ($P < 0.0001$, Wilcoxon rank-sum test; Extended Data Fig. 7e). As a control, preparations of formula were subjected to VLP purification and sequence analysis, which yielded no detections of viruses that infect animal cells (data not shown), indicating that these viruses were unlikely to originate as contamination from the formula products themselves.

To validate the metagenomic detections, VLP DNA and RNA samples were also tested by qPCR for their content of *Adenovirus*, *Torque teno virus*, *Enterovirus*, *Astroviridae*, *Sapovirus* and *Norovirus* sequences (Supplementary Table 4). The qPCR analysis also showed enrichment of viruses that infect human cells in the exclusively formula-fed cohort ($P = 0.0002$, Fisher's exact test; Fig. 3b).

Both populations studied above were from urban cohorts in the United States. To investigate whether our results hold more broadly, we analysed samples from a cohort of African newborns from Botswana. Infants who were 4 months of age were sampled using rectal swabs, so only qPCR assays and not sequence-based assays were carried out. All infants were delivered vaginally. We again found an association between exclusive formula feeding and colonization of viruses that grow in human cells ($P = 0.011$, Fisher's exact test; Fig. 3b).

Feeding type and other variables were then tested for effects on phage populations. Phage genes were annotated on assembled contigs, and their abundances used to calculate Bray–Curtis distances between communities. Significant differences in the population structure of phages could be detected based on feeding mode ($P = 0.001$, permutational multivariate analysis of variance (PERMANOVA); Extended Data Fig. 8a, c), but not for another 12 metadata variables (Extended Data Fig. 8a–e). To investigate the taxa involved, the shotgun sequencing analysis of whole stool from pooled discovery and validation cohorts was queried, which showed that whole-stool samples from breastfed infants contained a higher abundance of *Bifidobacterium* (false-discovery rate (FDR)-corrected $P = 0.02$; Fig. 3c) and *Lactobacillus* (FDR-corrected $P = 0.03$; Fig. 3c). Consistent with host abundances, VLP sequences that aligned to the temperate phages of *Bifidobacterium* and *Lactobacillus* were also enriched in breastfed infants ($P = 0.03$ and $P < 0.0001$, Fisher's exact test; Fig. 3d), in part explaining the effects of feeding mode on phage populations.

In summary, our data indicate that viral colonization in early life is stepwise, with the first phase characterized by the induction of prophages from pioneer bacteria, and a second phase involving colonization with viruses that infect human cells, the latter of which is modulated by breastfeeding (Fig. 3e). Previous epidemiological studies have emphasized the protective effects of breastfeeding in reducing viral gastroenteritis and infant death[9,10]. Mixed feeding of formula and breast milk is also reported to be protective compared with feeding of formula only[8], as was seen in the metagenomic analysis reported here. Factors in breast milk that are known to inhibit viral colonization include maternal antibodies, human milk oligosaccharides, lactoferrin and additional breast milk proteins[27–29]. The work reported here further develops our understanding of the protection conferred by breastfeeding in several respects. The metagenomic data (1) document the extent of subclinical infections with potentially pathogenic viruses; (2) highlight the potency of viral inhibition by breastfeeding; and (3) reveal the diversity of viruses affected, including viral families that cannot be grown in the laboratory and for which inhibition by breastfeeding is unstudied. In the African cohort, we not only found viruses that grow in human cells more commonly in exclusively formula-fed babies, but we also found more colonization in both feeding groups compared to US babies, emphasizing potential opportunities to intervene to reduce viral transmission to infants. The metagenomic methods described here will be useful for the assessment of the effects of different feeding strategies in diverse global settings to optimize the protection of infants from viral infections in the gut.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2192-1.

1. Breitbart, M. et al. Viral diversity and dynamics in an infant gut. *Res. Microbiol.* **159**, 367–373 (2008).
2. Lim, E. S. et al. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).
3. Liu, L. et al. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *Lancet* **388**, 3027–3035 (2016).
4. Oude Munnink, B. B. & van der Hoek, L. Viruses causing gastroenteritis: the known, the new and those beyond. *Viruses* **8**, 42 (2016).
5. Kim, M. S., Park, E. J., Roh, S. W. & Bae, J. W. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.* **77**, 8062–8070 (2011).
6. Lepage, P. et al. Dysbiosis in inflammatory bowel disease: a role for bacteriophages? *Gut* **57**, 424–425 (2008).
7. Hoyles, L. et al. Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Res. Microbiol.* **165**, 803–812 (2014).
8. Bahl, R. et al. Infant feeding patterns and risks of death and hospitalization in the first half of infancy: multicentre cohort study. *Bull. World Health Organ.* **83**, 418–426 (2005).
9. Arifeen, S. et al. Exclusive breastfeeding reduces acute respiratory infection and diarrhea deaths among infants in Dhaka slums. *Pediatrics* **108**, e67 (2001).
10. Victora, C. G. et al. Infant feeding and deaths due to diarrhea. A case–control study. *Am. J. Epidemiol.* **129**, 1032–1041 (1989).
11. Aagaard, K. et al. The placenta harbors a unique microbiome. *Sci. Transl. Med.* **6**, 237ra65 (2014).
12. Lauder, A. P. et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* **4**, 29 (2016).
13. Theis, K. R. et al. Does the human placenta delivered at term have a microbiota? Results of cultivation, quantitative real-time PCR, 16S rRNA gene sequencing, and metagenomics. *Am. J. Obstet. Gynecol.* **220**, 267.E1–267.E39 (2019).
14. de Goffau, M. C. et al. Human placenta has no microbiome but can contain potential pathogens. *Nature* **572**, 329–334 (2019).
15. Baumann-Dudenhoeffer, A. M., D'Souza, A. W., Tarr, P. I., Warner, B. B. & Dantas, G. Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes. *Nat. Med.* **24**, 1822–1829 (2018).
16. Reyes, A. et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
17. Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Hypervariable loci in the human gut virome. *Proc. Natl Acad. Sci. USA* **109**, 3962–3966 (2012).
18. Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10**, 607–617 (2012).
19. Aggarwala, V., Liang, G. & Bushman, F. D. Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mob. DNA* **8**, 12 (2017).
20. Wolf, Y. I. et al. Origins and evolution of the global RNA virome. *mBio* **9**, e02329-18 (2018).
21. Jacob, F., Sussman, R. & Monod, J. On the nature of the repressor ensuring the immunity of lysogenic bacteria (in French). *C. R. Acad. Sci.* **254**, 4214–4216 (1962).
22. Ptashne, M. *A Genetic Switch* (Blackwell, 1986).
23. Jacob, F. & Wollman, E. Spontaneous induction of the development of bacteriophage lambda during genetic recombination in *Escherichia coli K12* (in French). *C.R. Acad. Sci.* **239**, 317–319 (1954).
24. McNair, K., Bailey, B. A. & Edwards, R. A. PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* **28**, 614–618 (2012).
25. Dutilh, B. E. et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).
26. Shkoporov, A. N. et al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat. Commun.* **9**, 4781 (2018).
27. Turin, C. G. & Ochoa, T. J. The role of maternal breast milk in preventing infantile diarrhea in the developing world. *Curr. Trop. Med. Rep.* **1**, 97–105 (2014).
28. Newburg, D. S., Ruiz-Palacios, G. M. & Morrow, A. L. Human milk glycans protect infants against enteric pathogens. *Annu. Rev. Nutr.* **25**, 37–58 (2005).
29. Lewis, E. D., Richard, C., Larsen, B. M. & Field, C. J. The importance of human milk for immunity in preterm infants. *Clin. Perinatol.* **44**, 23–47 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Experimental model and human participants

Three cohorts of newborn infants were studied. Detailed information is provided in Supplementary Table 1. All experimentation complied with ethical regulations, and written informed consent was obtained from all human participants or their parents.

The Infant Growth and Microbiome Study (IGram) was approved by the Committee for the Protection of Human Subjects at The Children's Hospital of Philadelphia (IRB14-010833). African-American women planning to deliver at the Hospital of the University of Pennsylvania and their infants were enrolled. Inclusion and exclusion criteria are listed in Supplementary Table 5. Study visits were conducted at The Children's Hospital of Philadelphia. A total of 20 healthy, term infants were recruited for the discovery cohort. Stool samples were collected longitudinally at 0–4 days after birth (meconium samples, month 0), month 1 and month 4. The participants in an independent validation cohort had the same inclusion and exclusion criteria as the discovery cohort (only at month 4, $n = 86$). Fresh stool specimens from healthy infants were collected from diapers and aliquoted into faeces collection tubes (Sarstedt). All samples were stored at −80 °C. Metadata regarding delivery mode, infant feeding and health outcomes were collected by medical chart review and in-person interview by trained research personnel.

The Microbiome, Antibiotic and Growth Infant Cohort (MAGIC) Study was approved by the Committee for the Protection of Human Subjects at Children's Hospital of Philadelphia (IRB 15-012623). The study enrolled children born at Pennsylvania Hospital, receiving preventive healthcare in the CHOP Primary Care Network or participating in private practices, together with their biological mothers. The distribution of race, ethnicity and sex of the newborns reflected the general distribution in the participating sites. All infants enrolled were less than 120 h of age, greater than 36 weeks gestation, heavier than 2,000 g and spent fewer than 120 h in the neonatal care unit. Mothers were over the age of 18 and spoke English. A total of 39 heathy, term babies were used for this cohort. Study visits were conducted at Children's Hospital of Philadelphia. Stool samples were collected and questionnaires administered at birth and every 3 months until the infant reached 24 months of age. Stool samples obtained at 3 months of life were used for this cohort. Fresh stool specimens were collected at home by parents using a sterile faecal collection tube to scoop a pea-sized amount from a used diaper. Samples were then transported by courier on dry ice, aliquoted and stored at −80 °C. Maternal and baby clinical and metadata were collected by medical chart review and parent questionnaires.

The Botswana Infant Microbiome Study was approved by the Botswana Ministry of Health (IRB HPDME 13/8/1) and Institutional Review Boards at the University of Pennsylvania (IRB 822692) and Duke University (IRB 319561). Mother–infant pairs ($n = 300$) were enrolled within 48 h of delivery at the Princess Marina Hospital and two public clinics in or near Gaborone, Botswana. Exclusion criteria included maternal age less than 18 years, infant birth weight lower than 2,000 g, multiple gestation pregnancy and caesarian delivery. Participants were seen for monthly study visits until the infant was 6 months of age and every other month thereafter until the infant was 12 months of age. At all visits, a questionnaire was administered and clinical samples were obtained from the infant and the mother. Rectal swab samples obtained at 4 months of age from 100 infants were used for this cohort. These samples were collected into eNAT medium (Copan Italia), stored on ice and transported within 4 h to the National Health Laboratory in Gaborone for processing and storage at −80 °C. Metadata, including data regarding infant-feeding practices, were collected by medical chart review and in-person interview by trained research personnel.

In no cases were infants from any cohort reported to be suffering from gastroenteritis at the time of sampling.

Purification of VLPs from stool samples

VLPs were purified as previously described[30]. Approximately 200 mg of stool was homogenized in 30 ml of SM buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 8 mM $MgSO_4$), spun down and filtered through a 0.2-μm-pore-size filter (Thermo Fisher Scientific). The filtrate was concentrated using a 100-kDa-molecular-mass Amicon Ultra-15 Centrifugal Filter (Millipore), resuspended in 30 ml SM buffer and concentrated for the second time to a final volume of around 500 μl. The concentrate was treated with DNase I and RNase (Roche) at 37 °C for 30 min to degrade nonencapsulated nucleic acids. A total of 200 μl VLP preparation was used for viral nucleic acid extraction immediately after DNase I and RNase treatment; the remainder was stored at 4 °C for up to 3 months. To detect enveloped viruses, no chloroform was used to treat the VLPs. To test the purification efficiency of VLPs, 16S qPCR was used to quantify the 16S copy number before (total microbial DNA) and after purification of VLPs (VLP viral DNA). Samples showed an average reduction of 99.9% after purification (Extended Data Fig. 9).

Control spiking experiments with bacteriophage lambda showed that after addition to stool, approximately 90% of plaque-forming units could be recovered using the above methods.

VLP enumeration

Purified VLPs (35 μl) were diluted in 10 ml SM buffer and filtered onto a 0.02-μm Anodisc polycarbonate filter (Whatman). Filters were stained with 2× SYBR Gold (Thermo Fisher Scientific) for 15 min, then washed with H_2O once. After drying, the filter was mounted on a glass slide with 15 μl of mountant buffer (100 μl 10% ascorbic acid, 4.9 ml pH 7.4 PBS, 5 ml 100% glycerol; filtered through a 0.02-μm filter). For each filter, viruses were counted in 5–10 randomly selected fields of view. The filter was visualized using a motorized inverted system microscope IX81 (Shinjuku) for fluorescence. VLPs were counted using ImageJ. Stained particles of less than 0.5 μm in diameter were regarded as VLPs (larger particles were not counted). Purified lambda phages with known plaque-forming unit counts per ml were used as a positive control to adjust image colour, saturation, level and contrast. VLPs mock-purified from SM buffer were used as negative controls. At least one count per microscope field was set as a threshold for a positive detection, which was equal to around 6.6×10^6 counts per gram faeces. Lower than this threshold, the VLP counts were considered to be below the limit of detection. The results are listed in Supplementary Table 2.

Extraction and amplification of viral nucleic acids

Viral DNA and RNA were extracted from VLPs using the AllPrep DNA/RNA Mini kit (Qiagen) according to the manufacturer's instructions. DNA was stored at −20 °C and RNA at −80 °C. Viral DNA was subjected to DNA whole-genome amplification using the GenomiPhi V2 Amplification kit (GE Healthcare). Viral RNA was treated with DNase, reverse transcribed and PCR-amplified as previously described[31]. Specifically, 20 μl of RNA was treated with 10 units of RNase-free, recombinant DNase (Roche) for 20 min at 37 °C. A total of 5 μl of each sample was then reverse transcribed. First-strand cDNA synthesis was completed using the SuperScript III First Strand Synthesis kit (Thermo Fisher Scientific) and primer A (5′-GTTTCCCAGTCACGATCNNNNNNNNN-3′), to allow for random priming[31]. 'N' indicates a mixture of all four bases. Second-strand synthesis was performed using DNA Polymerase I, Large (Klenow) Fragment (New England Biolabs). The dsDNA product was then amplified by adding primer B (5′-GTTTCCCAGTCACGATC-3′) with AccuPrime Taq High Fidelity DNA polymerase (Thermo Fisher Scientific) with the following reaction conditions: 75.5 μl of molecular-grade H_2O, 10 μl of 10× PCR buffer I, 4 μl of 50 mM $MgCl_2$, 2.5 μl 10 mM dNTPs,

1 µl 100 µM primer B, 1 µl Taq and 6 µl dsDNA product. Products were amplified at 94 °C for 2 min, 94 °C for 30 s, 40 °C for 30 s, 50 °C for 30 s, 72 °C for 1 min for 40 cycles. Amplified DNA and cDNA were stored at −20 °C.

For the African cohort, samples were only available as rectal swabs stored in eNAT medium (Copan Italia), which contains guanidine thyocianate. It was therefore not possible to purify VLPs before analysis. On the basis of experience, such samples have such high human DNA content that shotgun metagenomic analysis yields overwhelmingly human sequences—we therefore carried out only qPCR analysis of these samples. Nucleic acids were purified using the AllPrep DNA/RNA Mini kit (Qiagen) as described above. No pre-amplification was performed for either DNA or RNA.

Total microbial DNA extraction

Approximately 200 mg of stool was used for total microbial DNA extraction. Total microbial DNA was purified from each sample using the Mo Bio PowerSoil kit (Mo Bio) following the manufacturer's instructions. A total of 50 µl total microbial DNA was obtained for each sample and stored at −20 °C.

Stool virome library and total microbial shotgun library construction and sequencing

Amplified viral DNA, cDNA and total microbial DNA were used for the construction of the shotgun libraries. The DNA concentration was measured using the Quant-iT PicoGreen dsDNA Assay kit (Thermo Fisher Scientific) and the fluorescence was detected by an EnVision Multilabel Plate Reader (Waltham). Libraries were made using an Illumina Nextera XT Samples Prep kit (Illumina), quantified using both the Quant-iT PicoGreen dsDNA Assay kit and the KAPA Library Quantification kit (Kapa Biosystems). The size distribution of the libraries was checked by 5300 Fragment Analyzer (Agilent). Libraries were pooled for sequencing. The concentration of the pooled libraries was measured using Qubit (Invitrogen) and the size distribution of the pooled libraries was checked by Agilent Technology 2100 Bioanalyzer using a High Sensitivity DNA chip (Agilent). Sequence was acquired using the Illumina Miseq (250-bp paired-end reads, Illumina) and HiSeq (150-bp paired-end reads, Illumina).

Isolation of bacterial strains

A total of 24 bacterial strains was isolated from the stool samples (19 samples from 12 infants) using three types of medium: Lysogeny broth medium in aerobic conditions, *Bifidus* selective medium (Sigma-Aldrich) and eosin methylene blue medium (Sigma-Aldrich) in anaerobic conditions, which were incubated at 37 °C for up to 72 h. Single colonies were picked and re-streaked in medium plates at least three times to isolate pure bacterial strains. The bacterial taxonomy was determined by matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI–TOF MS) using a MALDI–TOF BD instrument (BD) and default software. The taxonomy was further validated by mapping scaffolds to a 16S rRNA gene database.

Whole-genome sequencing of isolated bacterial strains

All 24 bacterial strains were cultured in Lysogeny broth overnight or until an optical density at 600 nm (OD_{600}) > 1. DNA was extracted using the phenol–chloroform method. DNA quality control was as described above. The TruSeq DNA PCR-Free kit (Illumina) was used to make genomic DNA-sequencing libraries; quality control and pooling was as described above. Sequencing was performed using the Illumina Miseq (250-bp paired-end reads, Illumina).

In vitro prophage induction and induced VLP sequencing

Overnight cultures of isolated bacterial strains were diluted 1/100 into 10 ml medium and grown until log phase (OD_{600} = 0.6). Mitomycin C (Sigma-Aldrich) was then added to a final concentration of 5 µg ml^{-1}. The OD_{600} values were measured and VLPs were purified after 6 h of culture. VLPs were purified from the bacterial culture using the same method as described for stool VLP purification without the homogenization step. The purified VLP DNA was extracted and amplified for virome sequencing as described above, and also enumerated by SYBR Gold staining. RNA phages are generally not thought to form prophages; RNA was interrogated for three samples, and no phages were identified.

In vitro phage infection

Overnight cultures of isolated bacterial strains were diluted 1/100 into 10 ml medium, grown to log phase (OD_{600} = 0.6), and then 100 µl bacteria was mixed with 100 µl of serial dilutions of isolated infant stool VLPs with the addition of $MgSO_4$ (with a final concentration of 10 nM). The mixture was incubated at 37 °C for 30 min, diluted in 3 ml of warm soft agar, and then plated on a pre-warmed Lysogeny broth plate. Lambda phage was used as a positive control.

16S rRNA gene qPCR

Bacterial abundance was quantified using qPCR of the V1–V2 region of the 16S rRNA gene using a TaqMan-based assay (Applied Biosystems). Primer, probe sequences and the PCR program were described previously[32] and are included in Supplementary Table 6. The reaction was conducted on a 7500 Fast Real Time qPCR system (Thermo Fisher Scientific). Triplicate reactions were performed. Results show the mean values (Supplementary Table 2). The limit of detection in the 16S qPCR assay was determined to be 20 copies per reaction, which was equal to around 2,000 copies per gram faeces.

Stool VLP sequence read quality control and taxonomic classification

Quality control for the stool VLP reads was performed using the Sunbeam pipeline[33] with a custom Sunbeam extension (https://github.com/guanxiangliang/sbx_dedup). In brief, low-quality reads and adaptor sequences were removed by Trimmomatic[34], low-complexity reads were identified and discarded by Komplexity (https://github.com/eclarke/komplexity) and then duplicate identical sequences (inferred PCR replicates) were filtered out by BBmap (https://jgi.doe.gov/data-and-tools/bbtools/). Dereplicated reads were aligned using BWA to the host (GRCh38 for human genome and GRCm38 for mouse genome) or phix174 and removed. The quality-controlled reads were classified by Kraken using a custom database that included all complete human, bacterial, archeal and viral genomes in RefSeq release 89 (released on 9 July 2018), with low-complexity regions masked before building the database.

To investigate environmental contamination or experimental reagent contamination, negative control samples were analysed, including empty diaper samples, empty stool container samples and reagent-only samples. The Decontam package in R was used on the Kraken classification data to remove contaminating species with 'prevalence' method[35] at a threshold of 0.5. Taxa including *Klebsiella* phage 0507-KN2-1, *Choristoneura occidentalis* granulovirus, *Vibrio* phage pYD38-A, *Pseudomonas* phage PPpW-4, *Burkholderia* phage ST79, *Burkholderia* phage KS9, *Bacillus* virus phi29, Simbu orthobunyavirus, and Shamonda orthobunyavirus were removed from downstream analysis.

Stool VLP sequence read assembly and annotation, and phage lifestyle prediction

The quality-controlled reads were assembled into contigs using MEGAHIT[36] for each individual. To quantify contigs in each sample, quality-controlled reads were mapped back to the contigs using Bowtie2[37], and the number of mapped reads was calculated by processing SAM files using custom code. To remove differences in sequencing depth, reads per million total reads (RPM) were calculated for each contig. Assembled contigs from virome libraries with length larger than 3,000 bp were selected to predict open-reading frames (ORFs)

using Prodigal in 'meta' mode[38]. The predicted ORFs were mapped to the viral protein database in UniProt Knowledgebase (TrEML and Swiss-Prot)[39] using BLASTP with $E < 10^{-5}$.

To exclude contigs resulting from contamination, we mapped negative control sample reads to the built VLP contigs. If the maximal RPM of negative control samples for the sample contig was greater than in the stool samples, then that contig was marked as contamination and not used for downstream analysis.

We defined an assembled contig as a viral contig if it had (1) at least one viral protein per 10 kb of VLP contig and (2) 50% of the predicted ORFs were viral ORFs. The taxonomy of each contig was classified as described previously[40] modified to compile attributions over multiple reading frames to generate a single taxonomic assignment. The ORFs were assigned to taxa based on the best-hit viral protein in the UniProt Knowledgebase. The majority taxonomic assignment over all ORFs within a contig was given to the contig. Contigs that could not be assigned to any taxa were classified as 'Others'. Contigs that were not assigned as 'Bacteriophage' were mapped to the NCBI nt database with a threshold of 80% coverage and 80% identity to futher remove contigs from non-viral genomes. In total, we identified 2,552 viral contigs among all 20 infants (Extended Data Fig. 2a, b). Contigs that shared the same taxonomic assignments were collapsed to yield pooled RPM values for each taxon. Viral richness was calculated by observed species number with RPM > 10. DNA virome reads that could be assigned to our set of viral contigs accounted for 11.3% ± 4.7% (mean ± s.e.m.) at month 0, 31.2% ± 5.6% at month 1 and 37.7% ± 5.4% at month 4 of all nonhuman reads (Extended Data Fig. 2c–e). Other reads come from contamination, genomes of other microorganisms and unassigned categories. We think that some of the unassiged reads represent unstudied bacteriophages, for which insufficient ORFs matched the viral ORFs in the database to label the contig as viral. For the RNA virome data, assembly from 12 out of 20 infants yielded contigs larger than 3,000 bp. The RNA virome reads that could be assigned to viral contigs accounted for a mean of 4.5% ± 4.7% (mean ± s.e.m.) at month 0, 15.9% ± 13.0% at month 1 and 10.0% ± 5.1% at month 4 of all nonhuman reads (Extended Data Fig. 2f–h).

Viral contigs were scored as temperate or lytic bacteriophages using PHACTS[24]. In order to obtain strong predictions, only viral contigs with at least 10 ORFs were analysed. Of 2552 viral contigs, 1029 were classified as "Bacteriophages" and contained more than 10 ORFs and used for the PHACTS analysis. Ten replicate PHACTS predictions were performed. Probability values obtained from PHACTS were standardized between -1 and 1, which was presented as probability of "Lytic" or "Temperate" (Extended Data Fig. 4b).

To test the abundance of crAssphages in the infant gut, we mapped the stool VLP reads to 37 genomes which belong to the crAssphage family[25,26,41,42]. At least 33% genome coverage was considered to be a positive detection (Extended Data Fig. 6). In this analysis, we included stool VLP sequencing data from a group of older healthy children (2–5 years old, $n = 21$; Supplementary Table 1).

Profiling human-cell viruses

Seven viral families that replicate on human cells were detected by Kraken. To further investigate the accumulation of these viruses, the viral genomes in RefSeq and Viral Neighbour databases that represent these families were retrieved from NCBI. The stool VLP sequences were mapped to these genomes to estimate genome coverage using Bowtie2 with global alignment option[37]. The output sam files were process by Samtools[43], Bedtools[44] and custom code (https://github.com/guanxiangliang/liang2019) to quantify the fraction of the genome covered. We favour use of percent coverage as a metric for genome detection[45]; amplification during sequence library preparation can yield many copies of single genome regions, yielding many sequence reads but with low genome coverage. Comparisons in several studies thus indicate coverage is a more reliable measure. We found that the negative control samples could contain coverage of up to ~10% of a viral genome (Extended Data Fig. 7k).

Human-cell virus qPCR

The numbers of selected human-cell viral genome copies in stool samples were determined by qPCR using TaqMan-based assays (Applied Biosystems). Primers and probes that target Adenoviruses[46], Human Torque teno viruses[47], Enteroviruses[48], Astroviruses[49], Sappovirus GI strains[50] and Norovirus GII strains[51] were used in this study. All primer and probe sequences are listed in Supplementary Table 6. The qPCR reactions were conducted on a 7500 Fast Real Time qPCR system using TaqMan Fast Advanced Master Mix (Thermo Fisher Scientific) in a final volume of 20 µl with 900 nM primers and 250 nM probe. All qPCR reactions were performed without preamplification for both VLP DNA or RNA. Triplicate reactions were performed and the results showing the mean values and standed deviations are listed in Supplementary Table 7.

The availability of metagenomic virome sequence data and qPCR data allowed assessment of qPCR efficiency given sporadic mismatches of viral sequences to qPCR primers. A comparision between virome sequencing and qPCR data are presented in Supplementary Table 4.

Total microbial shotgun metagenome sequencing read quality control and taxonomic classification

The quality control for the shotgun metagenome sequencing reads were performed using the default pipeline in Sunbeam. The quality-controlled reads were classified by Kraken using the same database as was used for stool VLP sequence analysis. To calculate the bacterial richness and diversity, 15,000 paired reads were randomly selected from each sample, and MetaPhlAn2 was used to align reads to different levels of bacterial taxonomy[52]. Bacterial richness was calculated as observed species number, and Shannon diversity was calculated using the Vegan package in R. The Decontam package in R was used to remove contaminating sequences with "prevalence" method at a threshold of 0.5[35].

Bacterial whole-genome sequence assembly and quality control

The quality control for the bacterial whole-genome sequence reads was performed using Sunbeam without removing low-complexity reads. The quality-controlled reads were assembled by SPAdes[53], followed by SSPACE to make scaffolds[54]. The quality of scaffolds (completeness and contamination) was evaluated by CheckM[55]. The assembled scaffolds revealed good quality for each bacterial strain (Supplementary Table 3).

Integrated analysis of stool VLP, induced VLP, shotgun metagenome and whole-genome sequence data

To analyse whether stool viruses from the stool VLP sequences matched sequences of induced VLPs from purified infant gut bacterial strains, reads from induced VLPs and stool VLPs were mapped to the corresponding bacterial scaffolds using Bowtie2[37]. The number of mapped reads and bedgraph files for coverage plots were generated using Samtools[43], Deeptools[44]and custom code. Induced VLP sequences from isolated bacterial strains were mapped to the stool virome contigs using the same method to assess whether the induced phages from isolated bacteria were more similar to stool VLP sequences from infants from whom the bacterial strain originated than in VLPs from unmatched individuals. The prophage genome annotation was performed by PHASTER[56] targeting bacterial genomic scaffolds longer than 100,000.

VLP contigs for analysis were identified as follows. First, we asked whether assembled contigs from the induced VLPs of the 24 purified bacterial strains could be identified in the 24 sequenced bacterial genomes. We required that more than 50% of the induced VLP contig length was matched to a bacterial genome scaffold or vice versa (Blastn with $E < 10^{-10}$). Second, we asked whether induced VLP contigs recognized as candidate prophages encoded proteins that were present in the UniProt viral protein database. At least 50% of the ORFs were

required to be virus-like proteins (Blastp with $E < 10^{-5}$). Third, for the induced VLPs, contigs were required to comprise at least 5% of all reads for inclusion in the analysis.

To evaluate whether the induced prophages from purified bacteria were more similar to stool VLP sequences from infants from which the bacterial strain originated than VLPs from unmatched infants, we mapped the stool VLP reads to the corresponding induced VLP contigs from the same infant (within infants) as well as unmatched infants (between infants) using Bowtie2 (Extended Data Fig. 5a).

Several further analyses were performed to investigate the correlation between the proportion of each bacterial species in the infant gut community and the proportion of prophages from that bacterial species in the infant gut virome.

Stool VLP sequences were mapped to the induced VLP contigs identified above using Bowtie2. The proportion of mapped reads from stool VLPs were divided by the total number of stool VLP reads to obtain the proportion, which represents the abundance of each bacterial prophage in the infant gut virome. The proportion of isolated bacteria in the infant gut community was represented by the proportion of shotgun reads that could be mapped to the isolated bacterial genome divided by all nonhuman reads. The abundance of bacterial prophages was plotted against the isolated bacterial abundance (Fig. 2d). This analysis was conducted using data based on both mitomycin C induction (Fig. 2d) and spontaneous induction (Extended Data Fig. 5b).

Phage population structure analysis

To analyse the phage populations, 185 samples pooled from both discovery and validation US cohorts were analysed. Assembled contigs from individual DNA virus libraries with length larger than 3,000 bp were selected to predict ORFs as described above. Accurately assigning taxonomic ranking of viral contig is still a challenge, therefore, we preformed a taxonomy-independent population analysis. ORFs were mapped to the Pfam database using HmmScan (HMMER 3.1; http://hmmer.org/) with $E < 10^{-5}$. Pfam entries that belong to phages, and those that were shared by phages and bacteria, were selected for further analysis. The coordinates of each Pfam entry on the contigs were identified by custom code, and VLP reads were aligned to these coordinates by featureCounts[57] to evaluate the abundance of each Pfam entry. The Pfam annotations for each sample were catalogued and a matrix was generated for annotation over all samples. Clustering was evaluated using Bray–Curtis dissimilarities. Bray–Curtis dissimilarities were plotted using principal coordinate analysis (PCoA), and differences among groups (infant age, infant feeding type, infant delivery type, infant gender, mother body type, formula type, mother pregnancy induce hypertension or diabetes and mother Chorioamnionitis) were tested using PERMANOVA. Continuous variables (gestational age, infant birth weight, household underage number, household number and mother pregnancy weight gain) were fit to the PCoA ordination by regression using the Envfit function. *P* values were determined using 999 permutations. The analysis was carried out using the vegan R package.

Bifidobacterium and *Lactobacillus* phage analysis

We downloaded 42 *Lactobacillus* phage genomes from RefSeq and used these for comparison. RefSeq did not contain any *Bifidobacterium* phage genome sequences, but two *Bifidobacterium* phage genomes were available in NCBI (accession numbers GQ141189.1 and MH444512.1) and were used for analysis here. Genome coverage was estimated using the same method as was used for animal virus coverage analysis. The *Bifidobacterium* and *Lactobacillus* phages that were highly covered (>33%) by sequencing contain annotated 'Integrase' proteins, suggesting temperate replication cycles.

Quantification and statistical analysis

Statistical tests were conducted using R. Nonparametric tests were used to compare two independent groups (Wilcoxon rank-sum test), two related groups (Wilcoxon signed-rank test) and multiple groups (Kruskal–Wallis test with Bonferroni correction). Nonparametric correlation was performed using Spearman's rank-order correlation (*R* represents Spearman's ρ). Fisher's exact test was used to test the difference between two categorical variables. *P* values for multiple comparisons were corrected using the Benjamini–Hochberg FDR method. $P < 0.05$ or FDR-corrected $P < 0.05$ was considered significant. All reported *P* values are from two-sided comparisons. All acquired data were included in analyses.

Gnotobiotic mouse control

As a control for this study, we prepared and analysed VLPs from stool samples from gnotobiotic mice and found viral sequences that were not present in contamination controls. In this case, the particles found were derived from murine endogeneous retroviruses, specifically murine leukaemia virus, which is known to be present in the germ line of the mouse strain C57BL/6 used here[58,59]. Evidently endogenous retroviral particles can be shed into the mouse gut and detected by our methods, providing a positive control for our analysis of human samples. No additional VLP contigs that passed quality filtering were detected.

Possible contribution of human endogeneous retroviruses

Human endogenous retroviruses (HERVs) are another candidate source of viral particles in neonates, and low levels of these sequences could be detected in VLP DNA fractions. However, HERV particles contain RNA, and HERVs were not detected significantly in RNA VLP fractions (Extended Data Fig. 10). Quality-control studies showed that HERV DNAs were probably contributed by contaminating human DNA, and were present in proportions predicted given the frequencies of other human genomic repeated sequences.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sample information and raw sequences are available in the National Center for Biotechnology Information Sequence Read Archive under BioProject ID PRJNA524703 (Supplementary Table 8). The isolated bacterial genome sequences have been deposited at DDBJ/ENA/GenBank under the accession numbers WVTF00000000–WVUC00000000 (Supplementary Table 3).

Code availability

All bioinformatic scripts are available on Github (https://github.com/guanxiangliang/liang2019).

30. Chehoud, C. et al. Transfer of viral communities between human individuals during fecal microbiota transplantation. *mBio* **7**, e00322-16 (2016).
31. Wang, D. et al. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.* **1**, e2 (2003).
32. Hill, D. A. et al. Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis. *Mucosal Immunol.* **3**, 148–158 (2010).
33. Clarke, E. L. et al. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome* **7**, 46 (2019).
34. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
35. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
36. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
37. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
38. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

39. Pundir, S., Magrane, M., Martin, M. J., O'Donovan, C. & The UniProt Consortium. Searching and navigating UniProt databases. *Curr. Protoc. Bioinformatics* **50**, 1.27.1–1.27.10 (2015).
40. Minot, S. et al. Rapid evolution of the human gut virome. *Proc. Natl Acad. Sci. USA* **110**, 12450–12455 (2013).
41. Yutin, N. et al. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38–46 (2018).
42. Guerin, E. et al. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* **24**, 653–664 (2018).
43. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
44. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
45. Abbas, A. A. et al. Bidirectional transfer of *Anelloviridae* lineages between graft and host during lung transplantation. *Am. J. Transplant.* **19**, 1086–1097 (2019).
46. Jothikumar, N. et al. Quantitative real-time PCR assays for detection of human adenoviruses and identification of serotypes 40 and 41. *Appl. Environ. Microbiol.* **71**, 3131–3136 (2005).
47. Abbas, A. A. et al. The perioperative lung transplant virome: torque teno viruses are elevated in donor lungs and show divergent dynamics in primary graft dysfunction. *Am. J. Transplant.* **17**, 1313–1324 (2017).
48. Verstrepen, W. A., Kuhn, S., Kockx, M. M., Van De Vyvere, M. E. & Mertens, A. H. Rapid detection of enterovirus RNA in cerebrospinal fluid specimens with a novel single-tube real-time reverse transcription-PCR assay. *J. Clin. Microbiol.* **39**, 4093–4096 (2001).
49. van Maarseveen, N. M., Wessels, E., de Brouwer, C. S., Vossen, A. C. & Claas, E. C. Diagnosis of viral gastroenteritis by simultaneous detection of adenovirus group F, astrovirus, rotavirus group A, norovirus genogroups I and II, and sapovirus in two internally controlled multiplex real-time PCR assays. *J. Clin. Virol.* **49**, 205–210 (2010).
50. Oka, T. et al. Detection of human sapovirus by real-time reverse transcription-polymerase chain reaction. *J. Med. Virol.* **78**, 1347–1353 (2006).
51. Rolfe, K. J. et al. An internally controlled, one-step, real-time RT-PCR assay for norovirus detection and genogrouping. *J. Clin. Virol.* **39**, 318–321 (2007).
52. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
53. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
54. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
55. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
56. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
57. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
58. Kao, D. et al. ERE database: a database of genomic maps and biological properties of endogenous retroviral elements in the C57BL/6J mouse genome. *Genomics* **100**, 157–161 (2012).
59. Young, G. R., Kassiotis, G. & Stoye, J. P. Emv2, the only endogenous ecotropic murine leukemia virus of C57BL/6J mice. *Retrovirology* **9**, 23 (2012).

Acknowledgements We thank J. You and co-workers for help with imaging, M. Goulian and co-workers for bacterial strains, and F. Rohwer and associates for assistance with phage induction; members of the Bushman laboratory for help and suggestions; and L. Zimmerman for artwork and help with the manuscript. The Botswana Infant Microbiome Study team thank Copan Italia for their donation of the eNAT medium and flocked swabs used for the collection of rectal swab specimens. This work was supported by NIH grants R61-HL137063 (F.D.B.), R01-HL113252 (F.D.B.) and R01DK107565 (G.D.W. and E.F.). The project described was also supported by the Penn Center for AIDS Research (P30 AI 045008) (F.D.B.), the PennCHOP Microbiome Program (F.D.B., G.D.W. and R.N.B.) and a Tobacco Formula grant under the Commonwealth Universal Research Enhancement (CURE) program (grant number SAP 4100068710) (F.D.B. and R.N.B.). Funding was also provided by an unrestricted donation from the American Beverage Foundation for a Healthy America to the Children's Hospital of Philadelphia to support the Healthy Weight Program (C.Z., K.B., E.F., J.S.G. and B.S.Z.). The project was also supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through grants UL1TR000003 and UL1TR001878 (E.F.). Funding for the Botswana Infant Microbiome Study was provided by the Duke Center for AIDS Research (5P30 AI064518) and the NIH (K23 AI135090).

Author contributions G.L. carried out biochemical analysis, sequencing and bioinformatic analysis; L.R.K., H.Z. and L.M. assisted with biochemical manipulations; C.Z., S.S.-M. and K.B. assisted with bioinformatic analysis; G.D.W., R.N.B., P.D., E.F., M.A.E., J.S.G., A.K., B.S.Z., M.S.K., M.Z.P. and T.M. carried out sample collection; G.L. and F.D.B. conceived the project and wrote the paper.

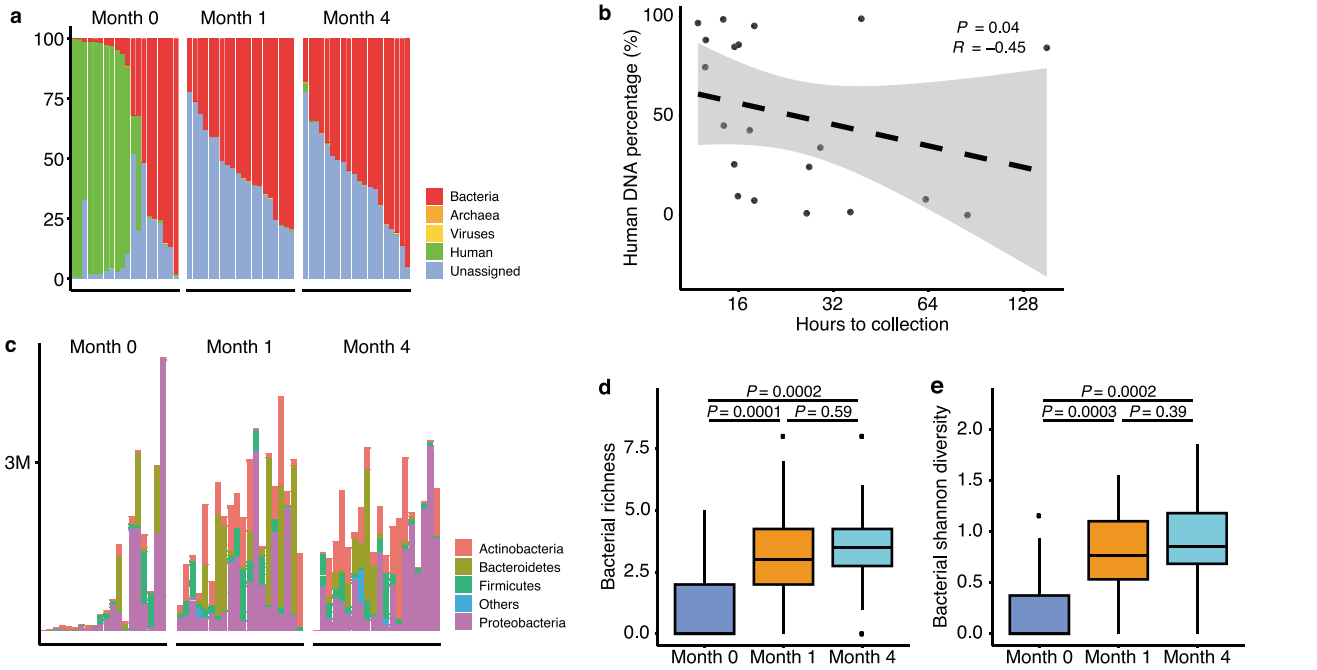
Competing interests The authors declare no competing interests.

Additional information

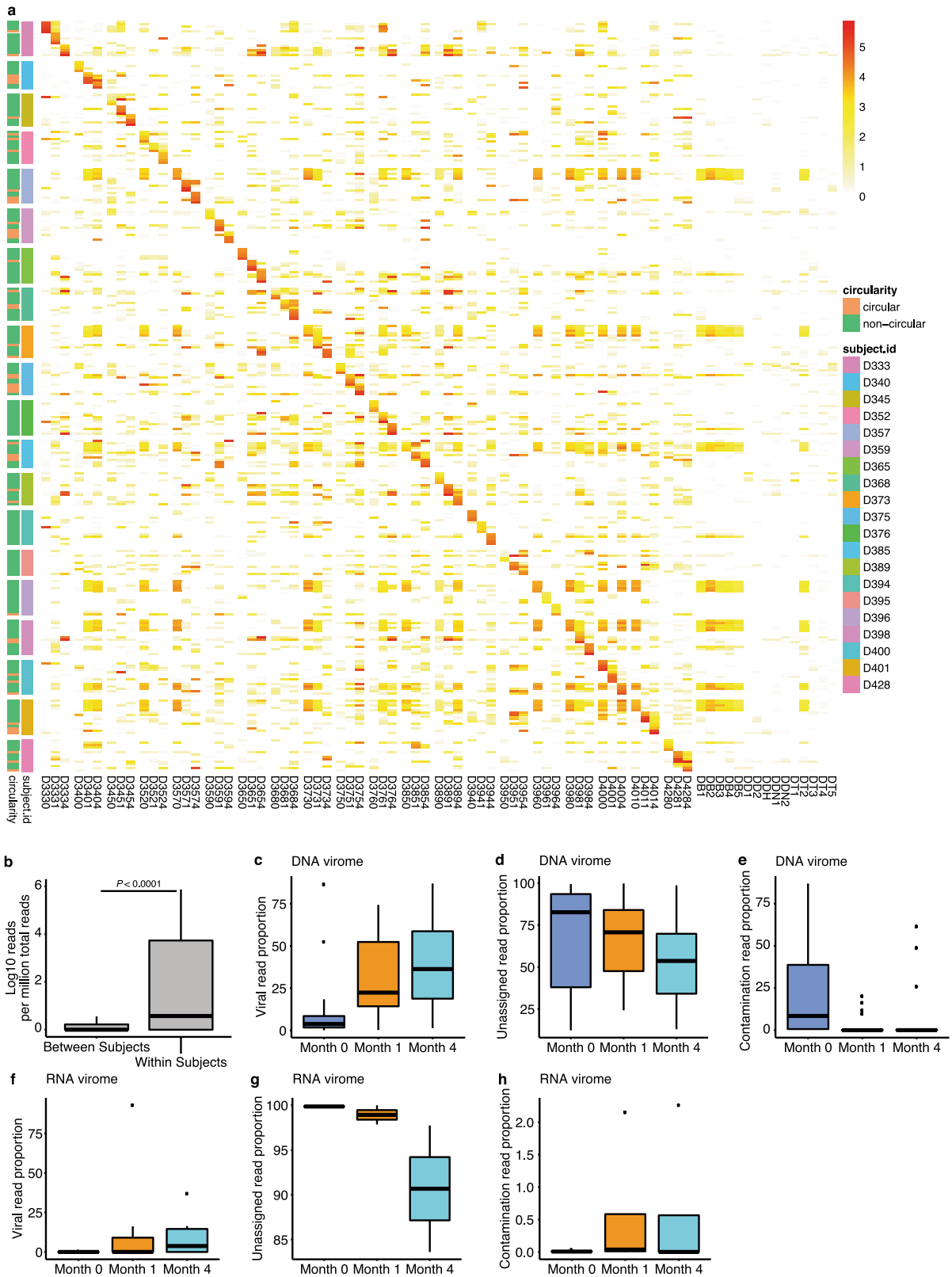
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2192-1.

Correspondence and requests for materials should be addressed to F.D.B.

Reprints and permissions information is available at http://www.nature.com/reprints.

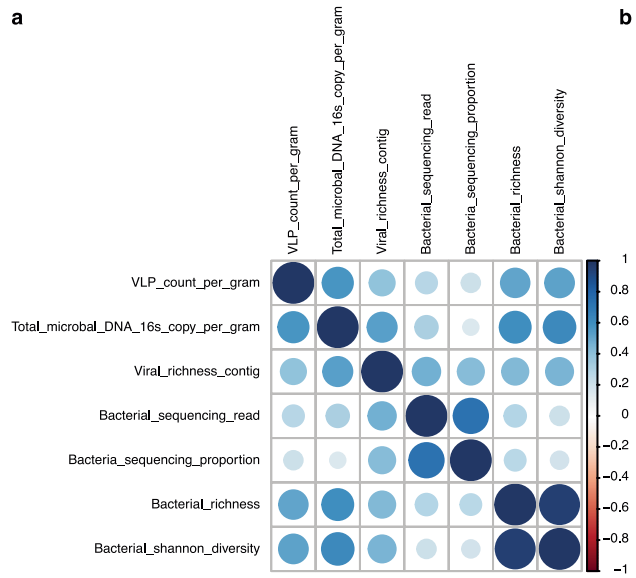


Extended Data Fig. 1 | Overview of total stool microbial shotgun metagenomic sequencing. a, Percentage of reads mapped to human or microbial genomes or that were unassigned. The types of DNA detected are indicated on the right. **b**, Correlation between the percentage of human DNA and sampling time after delivery using month-0 samples ($n = 20$). The percentage of human DNA is shown on the *y* axis, and the sampling time after delivery is shown on the *x* axis. The black dashed line shows the linear regression line and the grey-shaded region shows the 95% confidence interval for the slope. Two-sided Spearman's rank-order correlation method was used to test significance (R represents Spearman's ρ). **c**, Taxonomic composition of bacteria at the phylum level. The total read number is shown on the *y* axis; the *x* axis shows different samples. **d**, Bacterial richness. The *y* axis shows the richness calculated as the number of observed species. **e**, Bacterial diversity. **d**, **e**, A two-sided Wilcoxon rank-sum test was used to test the difference between different age groups ($n = 20$ infants at three time points). The horizontal lines in the box plots represent the third quartile, median and first quartile; whiskers extend to ±1.5× the interquartile range. The dots represent the outliers.



Extended Data Fig. 2 | Summary of virome sequencing of infant stool. a, Heat map summarizing the representation of the top five most-abundant DNA viral contigs in each sample. Samples are grouped sequentially by infant on both the *x* axis and *y* axis. The last group of infants on the *x* axis are negative control samples. Circularity indicates whether a contig is circular (orange colour) or not (light-green colour). The heat map map colour represents the abundance (log-transformed reads per million total reads) of each contig in each sample. **b**, Contig read abundance compared between different infants and within the same individuals. Time points were pooled for each individual. **c**–**e**, Percentage of DNA virome reads assigned to viruses (**c**), unassigned (**d**) and contamination (**e**). **f**–**h**, Percentage of RNA virome reads assigned to viruses (**f**), unassigned (**g**) and contamination (**h**). **b**–**h**, $n = 20$ infants at three time points were tested. The horizontal lines in box plots represent the third quartile, median and first quartile; whiskers extend to ±1.5× the interquartile range. The dots represent the outliers.

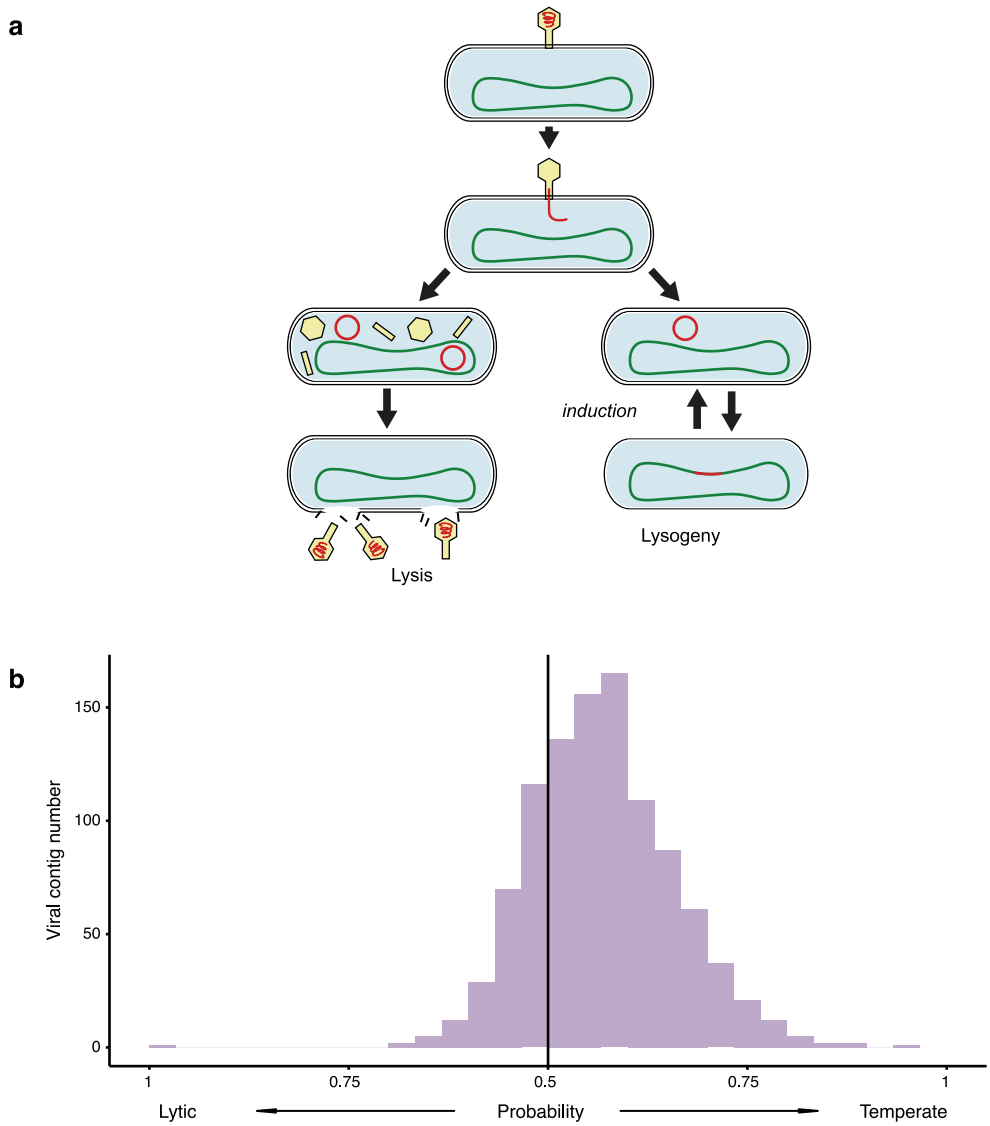
a



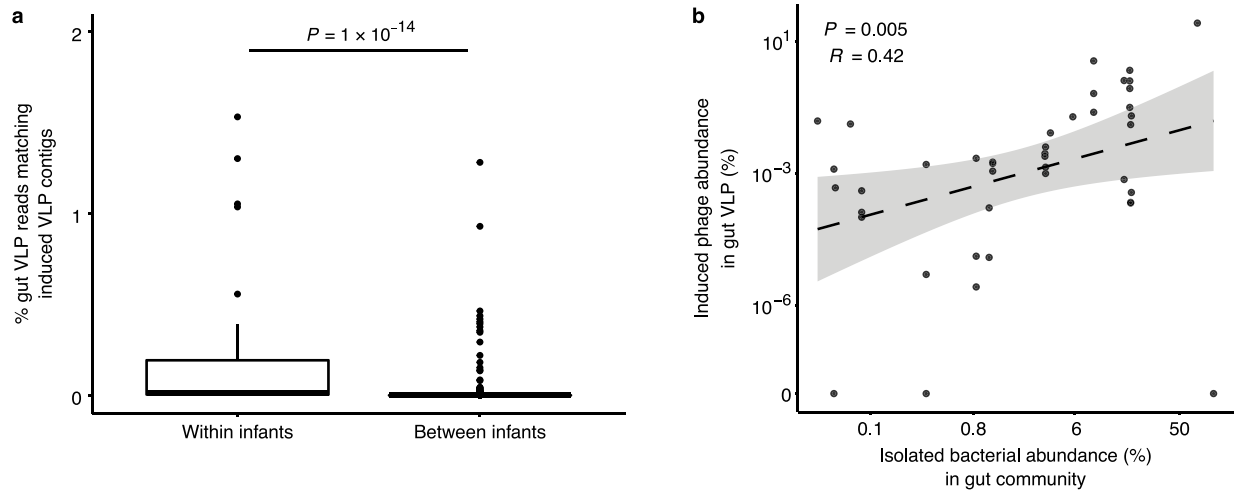
b

data1	data2	R.value	P.value	FDR
VLP_count_per_gram	VLP_count_per_gram	1.00	0.00e+00	0.00e+00
VLP_count_per_gram	Viral_richness_contig	0.40	1.68e−03	2.77e−03
VLP_count_per_gram	Total_microbal_DNA_16s_copy_per_gram	0.57	1.61e−06	5.68e−06
VLP_count_per_gram	Bacterial_shannon_diversity	0.52	1.92e−05	5.76e−05
VLP_count_per_gram	Bacterial_sequencing_read	0.28	3.00e−02	3.86e−02
VLP_count_per_gram	Bacterial_richness	0.51	2.98e−05	8.32e−05
VLP_count_per_gram	Bacteria_sequencing_proportion	0.22	9.81e−02	1.09e−01
Viral_richness_contig	VLP_count_per_gram	0.40	1.68e−03	2.77e−03
Viral_richness_contig	Viral_richness_contig	1.00	0.00e+00	0.00e+00
Viral_richness_contig	Total_microbal_DNA_16s_copy_per_gram	0.53	1.20e−05	3.88e−05
Viral_richness_contig	Bacterial_shannon_diversity	0.46	2.20e−04	4.81e−04
Viral_richness_contig	Bacterial_sequencing_read	0.47	1.40e−04	3.65e−04
Viral_richness_contig	Bacterial_richness	0.43	5.69e−04	1.07e−03
Viral_richness_contig	Bacteria_sequencing_proportion	0.43	6.06e−04	1.09e−03
Total_microbal_DNA_16s_copy_per_gram	VLP_count_per_gram	0.57	1.61e−06	5.68e−06
Total_microbal_DNA_16s_copy_per_gram	Viral_richness_contig	0.53	1.20e−05	3.88e−05
Total_microbal_DNA_16s_copy_per_gram	Total_microbal_DNA_16s_copy_per_gram	1.00	0.00e+00	0.00e+00
Total_microbal_DNA_16s_copy_per_gram	Bacterial_shannon_diversity	0.63	8.00e−08	3.41e−07
Total_microbal_DNA_16s_copy_per_gram	Bacterial_sequencing_read	0.32	1.28e−02	1.76e−02
Total_microbal_DNA_16s_copy_per_gram	Bacterial_richness	0.60	3.85e−07	1.48e−06
Total_microbal_DNA_16s_copy_per_gram	Bacteria_sequencing_proportion	0.15	2.51e−01	2.58e−01
Bacterial_shannon_diversity	VLP_count_per_gram	0.52	1.92e−05	5.76e−05
Bacterial_shannon_diversity	Viral_richness_contig	0.46	2.20e−04	4.81e−04
Bacterial_shannon_diversity	Total_microbal_DNA_16s_copy_per_gram	0.63	8.00e−08	3.41e−07
Bacterial_shannon_diversity	Bacterial_shannon_diversity	1.00	0.00e+00	0.00e+00
Bacterial_shannon_diversity	Bacterial_sequencing_read	0.22	9.42e−02	1.07e−01
Bacterial_shannon_diversity	Bacterial_richness	0.95	4.53e−32	2.82e−31
Bacterial_shannon_diversity	Bacteria_sequencing_proportion	0.19	1.40e−01	1.52e−01
Bacterial_sequencing_read	VLP_count_per_gram	0.28	3.00e−02	3.86e−02
Bacterial_sequencing_read	Viral_richness_contig	0.47	1.40e−04	3.65e−04
Bacterial_sequencing_read	Total_microbal_DNA_16s_copy_per_gram	0.32	1.28e−02	1.76e−02
Bacterial_sequencing_read	Bacterial_shannon_diversity	0.22	9.42e−02	1.07e−01
Bacterial_sequencing_read	Bacterial_sequencing_read	1.00	0.00e+00	0.00e+00
Bacterial_sequencing_read	Bacterial_richness	0.29	2.24e−02	2.98e−02
Bacterial_sequencing_read	Bacteria_sequencing_proportion	0.74	0.00e+00	0.00e+00
Bacterial_richness	VLP_count_per_gram	0.51	2.98e−05	8.32e−05
Bacterial_richness	Viral_richness_contig	0.43	5.69e−04	1.07e−03
Bacterial_richness	Total_microbal_DNA_16s_copy_per_gram	0.60	3.85e−07	1.48e−06
Bacterial_richness	Bacterial_shannon_diversity	0.95	4.53e−32	2.82e−31
Bacterial_richness	Bacterial_sequencing_read	0.29	2.24e−02	2.98e−02
Bacterial_richness	Bacterial_richness	1.00	0.00e+00	0.00e+00
Bacterial_richness	Bacteria_sequencing_proportion	0.27	3.42e−02	4.26e−02
Bacteria_sequencing_proportion	VLP_count_per_gram	0.22	9.81e−02	1.09e−01
Bacteria_sequencing_proportion	Viral_richness_contig	0.43	6.06e−04	1.09e−03
Bacteria_sequencing_proportion	Total_microbal_DNA_16s_copy_per_gram	0.15	2.51e−01	2.58e−01
Bacteria_sequencing_proportion	Bacterial_shannon_diversity	0.19	1.40e−01	1.52e−01
Bacteria_sequencing_proportion	Bacterial_sequencing_read	0.74	0.00e+00	0.00e+00
Bacteria_sequencing_proportion	Bacterial_richness	0.27	3.42e−02	4.26e−02
Bacteria_sequencing_proportion	Bacteria_sequencing_proportion	1.00	0.00e+00	0.00e+00

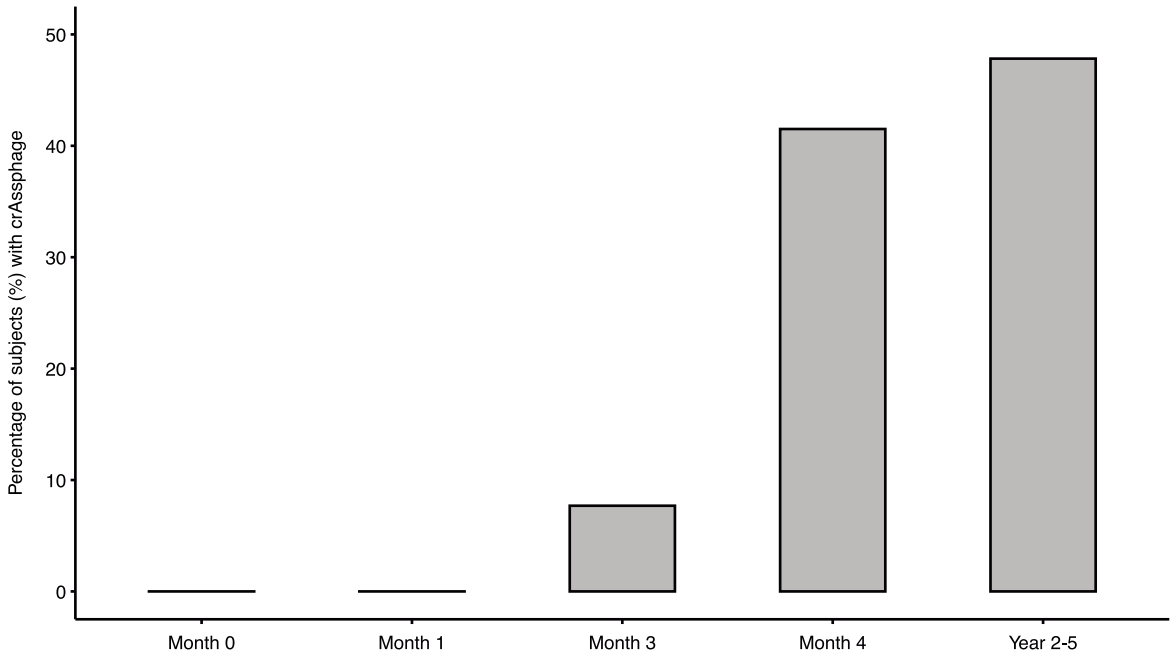
Extended Data Fig. 3 | Correlation between viral and bacterial communities. a, Pairwise correlations among sample measures including: VLP count number, bacterial 16S qPCR copy number, viral richness, bacterial sequence read proportion, bacteria richness and diversity. The size of circles indicates the *R* value of the correlation. Blue colour indicates a positive correlation, and red colour indicates a negative correlation. Samples from different time points were pooled ($n = 60$). A two-sided Spearman's rank-order correlation method was used in this analysis. **b**, As in **a**, but showing the raw data of the statistical analysis. *P* values, FDR-corrected *P* values and *R* (Spearman's ρ) values are shown.



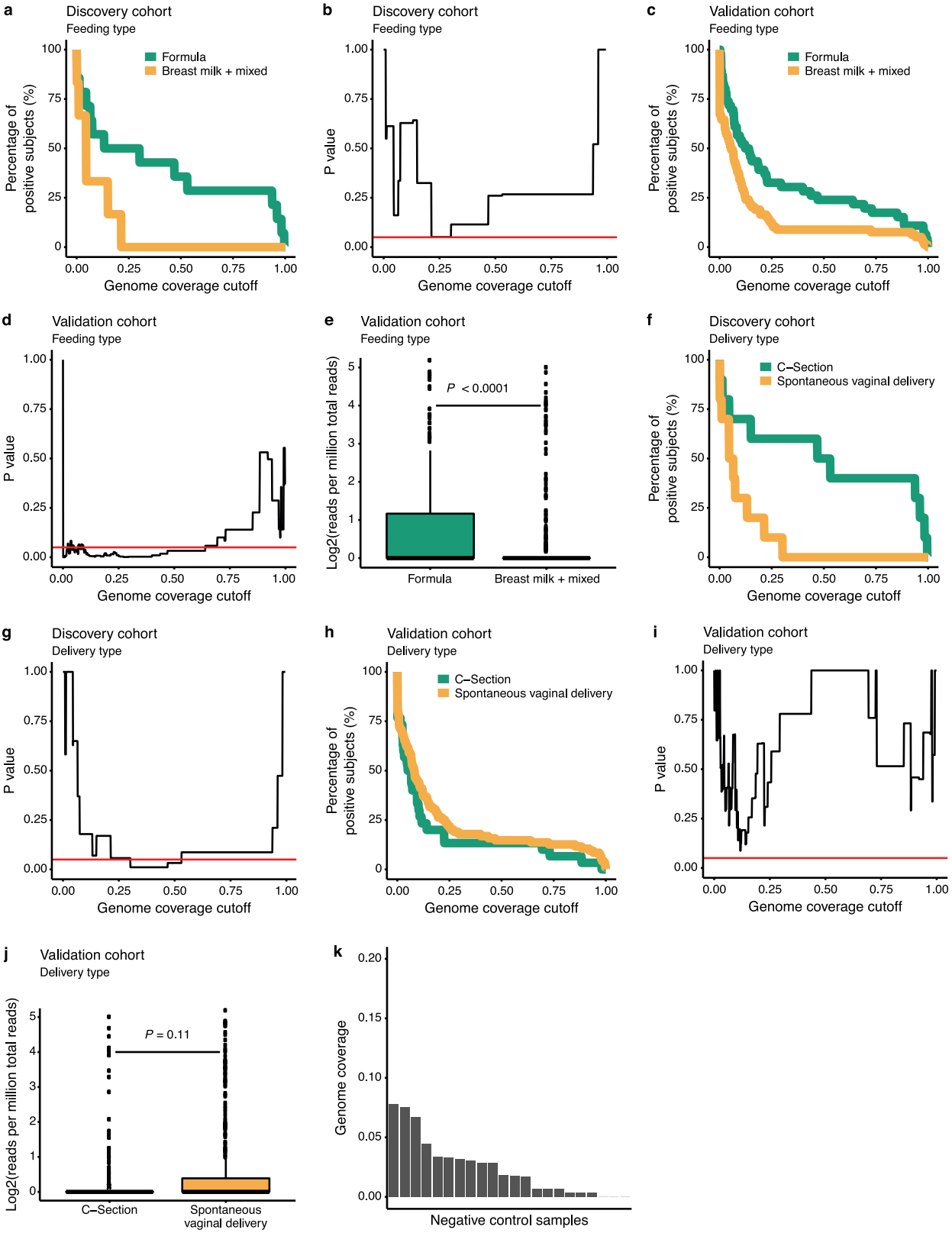
Extended Data Fig. 4 | Life cycles of bacteriophages. a, Diagram of lytic and lysogenic bacteriophage replication (based on a previous study[22]). Not shown are additional phage replication strategies, such as chronic infection and pseudolysogeny. **b**, Prediction of replication modes from contig sequences using PHACTS. The *x* axis shows the probability that a contig belongs to a lytic or temperate phage predicted by PHACTS. The *y* axis shows the viral contig number. In total, 1,029 phage contigs with at least 10 open-reading frames were used in this analysis. Of 1,029 contigs, 233 were predicted to be lytic and 794 were predicted to be temperate. Probability values obtained from PHACTS were standardized between −1 and 1, which was presented as a probability to be lytic or temperate.



Extended Data Fig. 5 | Prophage induction in the early-life virome. **a**, Comparison of the extent of sequence alignment of induced VLP sequences from bacterial strains compared with VLP sequences from stool samples. Contigs were generated from mitomycin-C-induced VLPs from purified bacterial strains from stool ($n = 33$ phage contigs from 16 bacterial isolates), then VLP reads from faeces were aligned to these contigs and quantified. 'Within infants' indicates matching stool VLPs to induced VLPs from purified bacteria for samples all from the same infant. 'Between infants' indicates alignment of stool VLPs versus induced VLPs from different infants. The horizontal lines in box plots represent the third quartile, median and first quartile. The dots represent the outliers. Samples were compared using a two-sided Wilcoxon rank-sum test. **b**, Correlation between the proportion of each bacterium in the infant gut community and the proportion of prophages from that bacterial species in the infant's gut virome. This plot is based on VLP sequences of phages produced by spontaneous induction ($n = 42$ phage contigs from 20 bacterial isolates). This is different from Fig. 2d, which is based on VLP sequences of phages produced after induction with mitomycin C. The black dashed line shows the linear regression line and the grey-shaded region shows the 95% confidence interval for the slope. The correlation was tested using a two-sided Spearman's rank-order correlation (R represents Spearman's ρ).



Extended Data Fig. 6 | Colonization by crAssphages in different age groups. The percentage of crAssphage-positive infants (as scored by requiring that the crAssphage genome was more than 33% covered by sequence reads from stool VLPs).

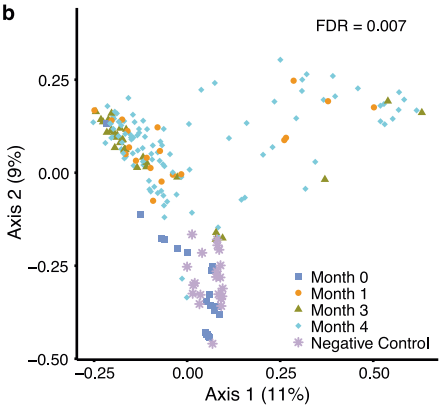


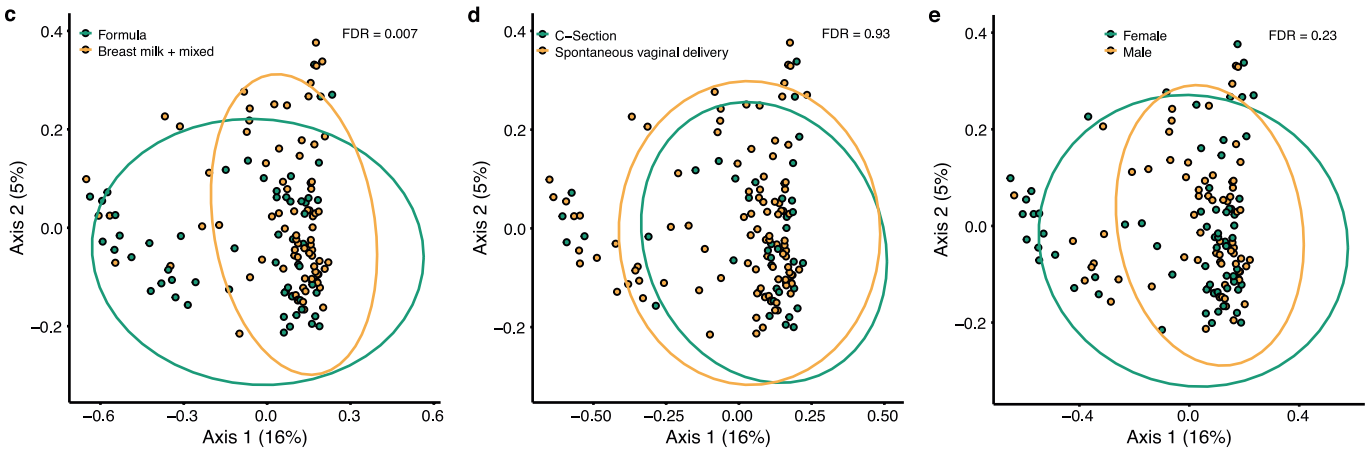
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Profiling of animal-cell viruses by virome sequencing. a, **c**, **f**, **h**, Percentage of infants positive for animal cell-associated viruses using different viral genome coverage cut-offs in the discovery cohort (**a**, **f**) and validation cohort (**c**, **h**). The green line shows the data from infants who were formula fed (**a**, **c**) or born by caesarean (C)-section delivery (**f**, **h**), and the yellow line shows the data from infants fed with breast milk or who were mixed fed (**a**, **c**) or were born by spontaneous vaginal delivery (**f**, **h**). **b**, **d**, **g**, **i**, Two-sided Fisher's exact test on infant feeding types (**b**, **d**) and delivery types (**g**, **i**) using different viral genome coverage cut-offs in the discovery cohort (**b**, **g**) and validation cohort (**d**, **i**). The horizontal red line indicates $P = 0.05$. **e**, **j**, Comparison of the relative abundance of animal-cell viruses between different feeding types (**e**) and delivery types (**j**). The abundance (reads per million total reads after log transformation) is shown on the *y* axis. A two-sided Wilcoxon rank-sum test was used to test the difference. The horizontal lines in box plots represent the third quartile, median and first quartile; whiskers extend to ±1.5× the interquartile range. The dots represent the outliers. **k**, Genome coverage fraction of negative control samples for animal-cell viruses. The maximal fraction of animal viral genome coverage for each negative control sample ($n = 25$) is shown on the *y* axis. Different negative control samples are shown on the *x* axis. Note that coverage never exceeds 10%. **a**, **b**, **f**, **g**, $n = 20$ samples from the discovery cohort were used; **c**–**e**, **h**–**j**, $n = 125$ samples from the validation cohort were used.

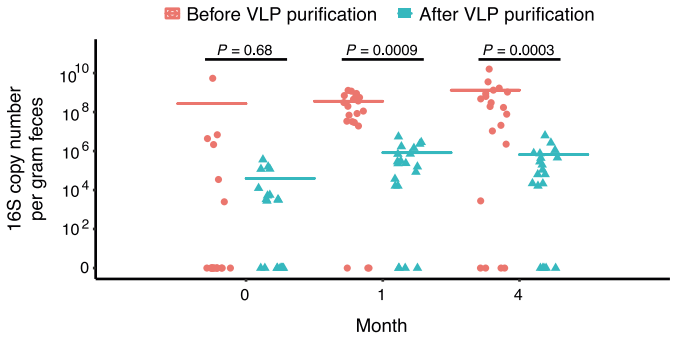
a

Variable	P	FDR
Infant age	0.001	0.007
Infant feeding type	0.001	0.007
Infant delivery type	0.870	0.930
Infant gender	0.050	0.233
Formula type	0.410	0.922
Mother body type	0.560	0.922
Mother pregnancy induced HTN or diabetes	0.590	0.922
Mother Chorioamnionitis	0.180	0.630
Gestational age	0.700	0.922
Birth weight	0.600	0.922
Mother maternal age	0.930	0.930
Mother pregnancy weight gain	0.340	0.922
Household underage number	0.750	0.922
Household number	0.790	0.922

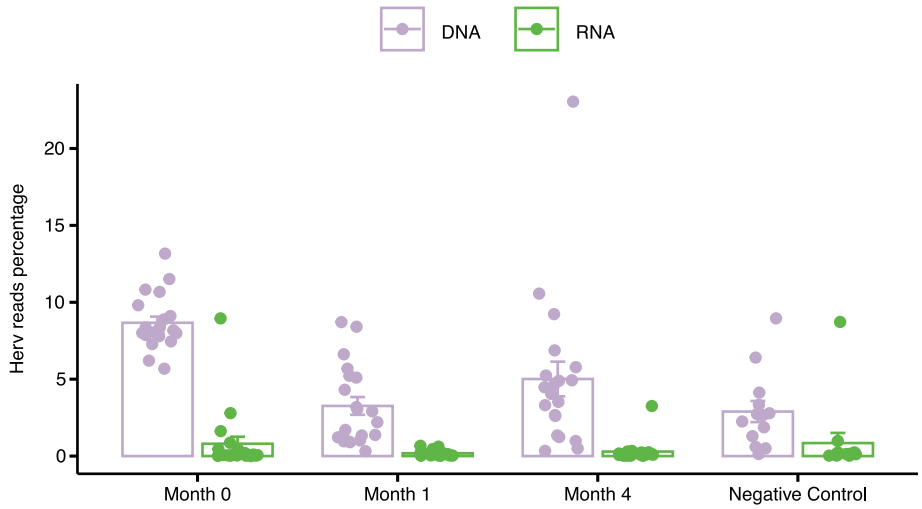




Extended Data Fig. 8 | Phage population structure. a, Statistical tests of the association of clinical variables with phage population structure. Variables are shown in the first column. *P* values and FDR-corrected *P* values are shown in the second and third columns. All categorized variables, such as infant age, infant feeding type, infant delivery type, infant gender, mother body type, formula type, mother pregnancy induced hypertension or diabetes and mother chorioamnionitis were tested by PERMANOVA. Continuous variables, including gestational age, infant birth weight, household underage number, household number and mother pregnancy weight gain were tested by Envfit. All samples from both discovery US and validation US cohorts ($n = 185$) were used to test infant age effects, and pooled samples at month 3 and month 4 from both discovery US and validation US cohorts ($n = 145$) were used to test other variables. **b**, PCoA plot based on phage Pfam counts per sample, coloured by infant ages. This analysis is based on the Bray–Curtis dissimilarity index for all stool samples from both discovery US and validation US cohorts ($n = 185$). Negative control samples were not included for Bray–Curtis dissimilarity assessment and statistical tests. **c**–**e**, PCoA plots of phage Pfam components, coloured by infant feeding types (**c**), delivery type (**d**) and infant gender (**e**). This analysis is based on pooled samples at month 3 and month 4 from both discovery US and validation US cohorts ($n = 145$), and as in **a**, PERMANOVA was used to test the differences. FDR-corrected *P* values are shown.



Extended Data Fig. 9 | 16S qPCR before and after VLP purification. Red and light-blue dots show before and after separately, and the horizontal lines represent the means ($n = 20$ infants at three time points were tested). A two-sided Wilcoxon signed-rank test was used to test the difference.



Extended Data Fig. 10 | Percentage of DNA aligning to sequences of HERVs in each sample. The percentage of HERV sequences in stool VLPs is shown on the *y* axis. Sample type and time point is shown on the *x* axis. The proportion of HERV sequences paralleled those of long interspersed nuclear elements and short interspersed nuclear elements, indicating that they are derived from human DNA contamination. Data are mean ± s.e.m.; $n = 20$ infants at three time points were tested.

nature research

Corresponding author(s): Frederic Bushman

Last updated by author(s): Jan 9, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Commercial softwares were not used for data collection.
Data analysis	qPCR was conducted and anazlyed on a 7500 Fast Real Time qPCR system. Raw sequence data was analyzed in Linux environment (Ubuntu 16.04.6 LTS). Softwares under Linux system included Sunbeam (v2.0),Trimmomatic (v0.36), Komplexity (0.3.0), BBmap (v38.22). Bowtie2 (v2.2.6), megahit (v1.1.3), prodigal (v2.6.3), Samtools (v1.7), HMMER (v3.1), Bedtools (v2.25), Blastp (v2.2.31), Blastn (v2.2.31), SPAdes (v3.12.0), SSPACE (v2.0), and checkM (v1.0.12). RStudio (v1.1.442) was used for downstream and statistical analysis. Softwares used in Rstudio included R (v3.4.4), MASS (v7.3-49), gtools (v3.8.1), viridis (v0.5.1), viridisLite (v0.3.0), gtable (v0.3.0), ggsci (v2.9), vegan (v2.5-5), lattice (v0.20-35), permute (v0.9-5), taxonomizr (v0.5.3), data.table (v1.12.2), qiimer (v0.9.4), randomcoloR (v1.1.0), RColorBrewer (v1.1-2), pheatmap (v1.0.12), tidyr (V0.8.3), reshape (v0.8.8), stringr (v1.4.0), dplyr (v0.8.3), plyr (v1.8.4), scales (v1.0.0), ggbeeswarm (v0.6.0), ggplot2 (v3.2.1), phyloseq (v1.22.3). MALDI Biotyper Realtime Classification and Biotyper software (v3.0) was used for MALDI-TOF BD instrument.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sample information and raw sequences are available in the National Center for Biotechnology Information Sequence Read Archive under BioProject ID

PRJNA524703. The isolated bacterial genome sequences have been deposited at DDBJ/ENA/GenBank under the accession WVTF00000000-WVUC00000000 (Supplementary Table 3). All bioinformatic scripts are available on Github (https://github.com/guanxiangliang/liang2019).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No initial sample size was calculated for our first cohort (20 infants) due to the exploratory nature of the study. We then sought to validate findings on the protective effect of breastfeeding, which did not achieve significance in the small initial cohort. We acquired a cohort that we calculated would likely provide sufficient resolution to yield a significant result given similar effect sizes. Assuming the breastmilk-exposed group had a proportion of 10% virus positive samples and our formula-only group had 40% virus positive samples (a conservative estimate from our discovery cohort), our collection of 79 breastmilk fed samples and 46 formula fed samples would provide a 96% chance of detecting a significant difference at P = 0.05. For our third cohort (African samples), we followed the power analysis used in assessing our second cohort.
Data exclusions	No available data was excluded.
Replication	For the analysis of the effects of breastfeeding, three separate cohorts were studied. Protective effect of breastfeeding was found in the discovery cohort, which did not achieve significance due to the small sample size. Two larger validation cohorts showed significant protective effects of breastfeeding.
Randomization	The observational cohort was segregated into formula-fed and breastmilk-exposed infants. Subjects were chosen to provide as balanced representation of the groups as possible given the cohort compositions. Subjects were subsampled randomly within each group where possible.
Blinding	Blinding was not possible in the context of analyzing breast feeding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Fecal samples were obtained as discarded animal biproducts from a centralized germ-free mouse facility (Protocol 805449; end date 8/28/2020). Species Mus musculus, strain C57BL/6 strain, male and female, 10 weeks of age.
Wild animals	No wild animals were used in this study.
Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	No ethical approval or guidance was required. Fecal samples were obtained as discarded animal biproducts from a centralized germ-free mouse facility (Protocol 805449; end date 8/28/2020) in the University of Pennsylvania.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	The Infant Growth and Microbiome Study (IGram) was approved by the Committee for the Protection of Human Subjects at The Children's Hospital of Philadelphia (IRB14-010833). African-American women planning to deliver at the Hospital of the University of Pennsylvania and their infants were enrolled. Study visits were conducted at The Children's Hospital of Philadelphia. A total of 20 healthy, term infants were recruited for the discovery cohort. Stool samples were collected longitudinally at day 0 to 4 days after birth (meconium samples, Month 0), month 1 (Month 1), and month 4 (Month 4). The participants in an independent validation cohort had the same inclusion and exclusion criteria as the discovery cohort (only at month 4, n = 86). Metadata regarding delivery mode, infant feeding and health outcomes was collected by medical chart review and in-person interview by trained research personnel. The Microbiome, Antibiotic, and Growth Infant Cohort (MAGIC) Study was approved by the Committee for the Protection of Human Subjects at Children's Hospital of Philadelphia (IRB 15-012623). The study enrolled children born at Pennsylvania Hospital, Philadelphia, PA, receiving preventive health care in the CHOP Primary Care Network or participating in private practices, together with their biological mothers. The distribution of race, ethnicity, and sex of the newborns reflected the general distribution in the participating sites. All subjects enrolled were less than 120 hours of age, greater than 36 weeks gestation, greater than 2000 grams, and spent less than 120 hours in the neonatal care unit. Mothers were over the age of 18 and spoke English. A total of 39 heathy, term babies were used for this cohort. Study visits were conducted at Children's Hospital of Philadelphia. Stool samples were collected and questionnaires administered at birth and every 3 months until the subject reached 24 months of age. Stool samples obtained at 3 months of life were used for this cohort. Mother and baby clinical and metadata were collected via medical chart review and parent questionnaires. The Botswana Infant Microbiome Study was approved by the Botswana Ministry of Health (IRB HPDME 13/8/1) and Institutional Review Boards at the University of Pennsylvania (IRB 822692) and Duke University (IRB 319561). Mother-infant pairs (n = 300) were enrolled within 48 hours of delivery at Princess Marina Hospital and two public clinics in or near Gaborone, Botswana. Exclusion criteria included maternal age less than 18 years, infant birth weight less than 2000 grams, multiple gestation pregnancy, and Caesarian delivery. Participants were seen for monthly study visits until the infant was 6 months of age and every other month thereafter until the infant was 12 months of age. At all visits, a questionnaire was administered and clinical samples were obtained from the infant and the mother. Metadata, including data regarding infant feeding practices, were collected by medical chart review and in-person interview by trained research personnel.
Recruitment	The Infant Growth and Microbiome Study (IGram) recruited African-infant born at the Hospital of the University of Pennsylvania. Study visits were conducted at The Children's Hospital of Philadelphia. The Microbiome, Antibiotic, and Growth Infant Cohort (MAGIC) enrolled children born at Pennsylvania Hospital, Philadelphia, PA, receiving preventive health care in the CHOP Primary Care Network or participating in private practices, together with their biological mothers. The distribution of race, ethnicity, and sex of the newborns reflected the general distribution in the participating sites. All subjects enrolled were less than 120 hours of age, greater than 36 weeks gestation, greater than 2000 grams, and spent less than 120 hours in the neonatal care unit. Mothers were over the age of 18 and spoke English. The cohorts above are US/urban cohorts which may not perfectly reflect acquisition of viruses in other settings, therefore, we included another cohort from developing country. The Botswana Infant Microbiome Study recruited mother-infant pairs within 48 hours of delivery at Princess Marina Hospital and two public clinics in or near Gaborone, Botswana. Exclusion criteria included maternal age less than 18 years, infant birth weight less than 2000 grams, multiple gestation pregnancy, and Caesarian delivery. Participants were seen for monthly study visits until the infant was 6 months of age and every other month thereafter until the infant was 12 months of age.
Ethics oversight	The Infant Growth and Microbiome Study (IGram) was approved by the Committee for the Protection of Human Subjects at The Children's Hospital of Philadelphia (IRB14-010833). The Microbiome, Antibiotic, and Growth Infant Cohort (MAGIC) Study was approved by the Committee for the Protection of Human Subjects at Children's Hospital of Philadelphia (IRB 15-012623). The Botswana Infant Microbiome Study was approved by the Botswana Ministry of Health (IRB HPDME 13/8/1) and Institutional Review Boards at the University of Pennsylvania (IRB 822692) and Duke University (IRB 319561).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Bacterial metabolism of bile acids promotes generation of peripheral regulatory T cells

https://doi.org/10.1038/s41586-020-2193-0

Received: 15 June 2019

Accepted: 14 February 2020

Published online: 15 April 2020

Check for updates

Clarissa Campbell[1,10]✉, Peter T. McKenney[1,2,10], Daniel Konstantinovsky[3], Olga I. Isaeva[4,5], Michail Schizas[1], Jacob Verter[1], Cheryl Mai[6], Wen-Bing Jin[7], Chun-Jun Guo[7], Sara Violante[8], Ruben J. Ramos[8], Justin R. Cross[8], Krishna Kadaveru[2], John Hambor[2] & Alexander Y. Rudensky[1,5,7,9]✉

Intestinal health relies on the immunosuppressive activity of CD4$^+$ regulatory T (T_{reg}) cells[1]. Expression of the transcription factor Foxp3 defines this lineage, and can be induced extrathymically by dietary or commensal-derived antigens in a process assisted by a *Foxp3* enhancer known as conserved non-coding sequence 1 (CNS1)[2–4]. Products of microbial fermentation including butyrate facilitate the generation of peripherally induced T_{reg} (pT_{reg}) cells[5–7], indicating that metabolites shape the composition of the colonic immune cell population. In addition to dietary components, bacteria modify host-derived molecules, generating a number of biologically active substances. This is epitomized by the bacterial transformation of bile acids, which creates a complex pool of steroids[8] with a range of physiological functions[9]. Here we screened the major species of deconjugated bile acids for their ability to potentiate the differentiation of pT_{reg} cells. We found that the secondary bile acid 3β-hydroxydeoxycholic acid (isoDCA) increased Foxp3 induction by acting on dendritic cells (DCs) to diminish their immunostimulatory properties. Ablating one receptor, the farnesoid X receptor, in DCs enhanced the generation of T_{reg} cells and imposed a transcriptional profile similar to that induced by isoDCA, suggesting an interaction between this bile acid and nuclear receptor. To investigate isoDCA in vivo, we took a synthetic biology approach and designed minimal microbial consortia containing engineered *Bacteroides* strains. IsoDCA-producing consortia increased the number of colonic RORγt-expressing T_{reg} cells in a CNS1-dependent manner, suggesting enhanced extrathymic differentiation.

Bile acids are cholesterol-derived molecules that are involved in essential physiological processes, including nutrient absorption, glucose homeostasis and regulation of energy expenditure[10]. Upon feeding, endocrine signals stimulate the emptying of the gallbladder into the duodenum, where bile acids aid in the emulsification of dietary fats[10]. Primary or liver-derived bile acids in mammals are mostly conjugated with taurine or glycine, and undergo pervasive deconjugation by microbial bile salt hydrolases in the small intestine. Although most bile acids are transported back into the liver via enterohepatic circulation, a small fraction of this pool (roughly 5%) escapes reabsorption in the ileum and is subject to further bacterial transformation in the colon, giving rise to secondary bile acids[10].

Knowing that the generation of colonic pT_{reg} cells is affected by microbial metabolites, we screened the major species of deconjugated bile acids found in mice and humans for their ability to enhance Foxp3 induction in vitro (Fig. 1a–c). Two secondary bile acids—ω-muricholic acid (ω-MCA) and isoDCA—potently increased the frequency of Foxp3$^+$ cells among naive CD4$^+$ T cells stimulated in the presence of DCs under suboptimal T_{reg} cell-inducing conditions (Fig. 1d). The differentiation of pro-inflammatory T_H17 cells was not affected by the presence of either bile acid (Extended Data Fig. 1a), suggesting a specific effect on the generation of T_{reg} cells.

Because both isoDCA and ω-MCA are isomers of molecules with no T_{reg} cell-promoting activity, we hypothesized that the spatial orientation of specific hydroxyl (-OH) groups is required for their effects. The formation of isoDCA from DCA requires oxidation of the 3α-OH group to an -oxo intermediate and its subsequent reduction into a 3β-OH group[11]. Despite remaining poorly characterized, the conversion of β-MCA into ω-MCA has also been reported to generate an -oxo intermediate[12,13]. We observed that the 3-oxo-derivative of DCA failed to potentiate T_{reg} cell induction to the same extent as isoDCA (Fig. 1e). Similarly, oxidation of the 6α-OH group abolished T_{reg} cell induction by

[1]Immunology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]SHINE Program, Research Beyond Borders, Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT, USA. [3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [4]Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia. [5]BostonGene LLC, Lincoln, MA, USA. [6]Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD–PhD Program, New York, NY, USA. [7]Immunology and Microbial Pathogenesis Program, Weill Cornell Medical College, Cornell University, New York, NY, USA. [8]Donald B. and Catherine C. Marron Cancer Metabolism Center, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [9]Howard Hughes Medical Institute and Ludwig Center, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [10]These authors contributed equally: Clarissa Campbell, Peter T. McKenney. ✉e-mail: campbec2@mskcc.org; rudenska@mskcc.org

a

Bile acid	R1	R2	R3	R4
Chenodeoxycholic acid (CDCA)	(α)OH	H	(α)OH	H
Cholic acid (CA)	(α)OH	H	(α)OH	(α)OH
Lithocholic acid (LCA)	(α)OH	H	H	H
Deoxycholic acid (DCA)	(α)OH	H	H	(α)OH
Isolithocholic acid (isoLCA)	(β)OH	H	H	H
Isodeoxycholic acid (isoDCA)	(β)OH	H	H	(α)OH
Ursodeoxycholic acid (UDCA)	(α)OH	H	(β)OH	H
β-Muricholic acid (β-MCA)	(α)OH	(β)OH	(β)OH	H
α-Muricholic acid (α-MCA)	(α)OH	(β)OH	(α)OH	H
γ-Muricholic acid (γ-MCA)	(α)OH	(α)OH	(α)OH	H
ω-Muricholic acid (ω-MCA)	(α)OH	(α)OH	(β)OH	H

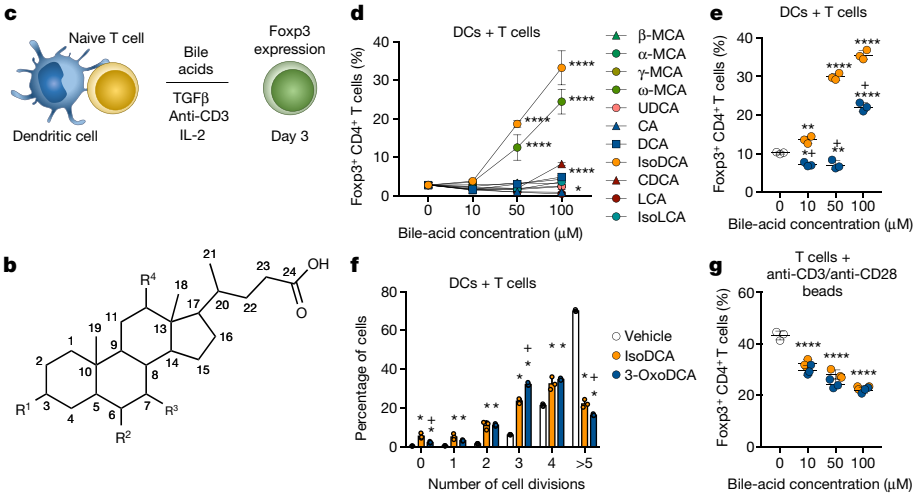


Fig. 1 | Bacterial epimerization of bile acids generates molecules with T_{reg} cell-inducing activity. a, **b**, Types of bile acids (**a**), their basic structure (**b**), and a summary of substitutions around the cholesterol backbone (**a**). **c**, Screen setup: naive $CD4^+$ T cells (5×10^4) were cocultured with DCs (1×10^5) in suboptimal T_{reg} cell-inducing conditions (1 ng ml^{-1} transforming growth factor (TGF)-β, 1 μg ml^{-1} CD3 antibody, 100 U ml^{-1} IL-2) and analysed on day 3 by fluorescence-activated cell sorting (FACS). The bile acids listed in **a** were added at the doses indicated in **d**, **e**, **g**. **d**, Frequencies of $Foxp3^+$ $CD4^+$ T cells after exposure to various concentrations of bile acids. **e**, The 3β-OH group of isoDCA is required for its T_{reg} cell-inducing effects. Cells were cocultured as described in **c** and incubated with 3-oxoDCA or isoDCA at the indicated concentrations. **f**, Assessment of cell proliferation. Naive $CD4^+$ T cells were labelled with Cell Trace Violet (CTV) and cultured with DCs as in **c**, in the presence of isoDCA or 3-oxoDCA (100 μM). CTV dilution was assessed on day 3 by FACS. The *y*-axis shows the percentage of cells that underwent the indicated number of cell divisions. **g**, Effect of bile acids on T_{reg} cell differentiation in the absence of DCs. Naive $CD4^+$ T cells were activated with CD3/CD28 antibody-coated beads under suboptimal T_{reg} cell-inducing conditions. IsoDCA and 3-oxoDCA were added at the indicated concentrations and cells were analysed on day 3 by FACS. Shown are means ± s.d. of technical replicates (**d**, $n = 4$; **e**–**g**, $n = 3$). $^*P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$ versus vehicle; plus symbol, $P < 0.0001$ versus isoDCA (paired concentration) by one-way (**d**, **e**, **g**) or two-way (**f**) analysis of variance (ANOVA) followed by a Dunnett (**d**) or Tukey's (**e**–**g**) multiple comparison test. Data are representative of at least three independent experiments.

ω-MCA (Extended Data Fig. 1b), indicating that microbial epimerization of bile acids gives rise to metabolites with unique immunomodulatory properties. Unlike ω-MCA, isoDCA is found at substantial levels (roughly 50 μM)[14] in the intestinal contents of healthy adult humans, and its biosynthesis is well characterized[11]. Therefore, we focused on understanding the biological activity of isoDCA.

The amphipathic nature of bile acids makes these molecules natural detergents[15] with potentially deleterious effects on cell viability and proliferation. Given that conditions associated with reduced T-cell proliferation are conducive to Foxp3 expression[16,17], we assessed whether this mechanism can account for the potentiation of bile acid-mediated T_{reg} cell differentiation. IsoDCA significantly reduced T cell proliferation compared with vehicle (Fig. 1f). We observed a similar effect in cells treated with 3-oxoDCA, which did not promote T_{reg} cell generation in a dose-dependent manner, indicating that decreased proliferation alone cannot account for the increased frequency of $Foxp3^+$ T cells. We next assessed whether DCs were required for isoDCA-mediated potentiation of T_{reg} cell induction. Neither isoDCA nor 3-oxoDCA increased T_{reg} cell frequencies when naive T cells were activated by beads coated with CD3/CD28 antibodies. Rather, both bile acids caused a decrease in the percentage of $Foxp3^+$ cells (Fig. 1g), which was also accompanied by reduced proliferation (Extended Data Fig. 1c). Consistent with these observations, isoDCA still enhanced T_{reg} cell frequencies when naive T cells lacking the *Foxp3* enhancer CNS3 were cocultured with DCs (data not shown). Together, these results support a requirement for antigen-presenting cells (APCs) in mediating the T_{reg} cell-inducing effects of isoDCA, and suggest a T cell-extrinsic immunoregulatory mechanism that is distinct from those reported for other bile acids[18,19].

Bile acids affect mammalian physiology through interactions with numerous receptors, including the farnesoid X receptor (FXR)[9]. We therefore used naive $CD4^+$ T cells from *CD4*^{cre} *Nr1h4*$^{fl/fl}$ and *CD4*^{WT} *Nr1h4*$^{fl/fl}$ mice as well as DCs from *Csf1r*^{cre} *Nr1h4*$^{fl/fl}$ and *Csf1r*^{WT} *Nr1h4*$^{fl/fl}$ mice to assess a potential role for FXR in the induction of T_{reg} cells by isoDCA in vitro. FXR deficiency in T cells did not change Foxp3 induction in response to isoDCA (Fig. 2a, left panel). By contrast, DCs lacking FXR generated a higher frequency of $Foxp3^+$ cells at baseline, and this increase could not be further enhanced by addition of isoDCA (Fig. 2a, right panel). These results support our finding that isoDCA acts upon DCs to potentiate the induction of T_{reg} cells, and suggest that FXR is involved in this process. To rule out effects from potential differences in the splenic DC population used as APCs, we analysed their composition in *Csf1r*^{cre} *Nr1h4*$^{fl/fl}$ mice and wild-type littermate controls. Although we failed to detect differences in the composition of the $CD11c^+$ MHC class II^+ cell population (mostly DCs) in the spleen and in other organs (Extended Data Fig. 2b), we did observe higher numbers of $Foxp3^+$ cells in the large intestine lamina propria (LILP) of *Csf1r*^{cre} *Nr1h4*$^{fl/fl}$ mice (Extended Data Fig. 2e), particularly of the RORγt$^+$ $Foxp3^+$ subset (Extended Data Fig. 2e). As pT_{reg} cells that arise in response to microbial antigens are predominantly RORγt$^+$ (refs. [20,21]), these data suggest that the absence of FXR in the myeloid compartment facilitates extrathymic generation of T_{reg} cells in vivo.

We next carried out RNA-sequencing (RNA-seq) analysis to comprehensively assess the effects of isoDCA treatment and FXR ablation in DCs. DCs exposed to isoDCA showed reduced expression of several genes related to antigen processing and presentation, including *Ciita*, *Ctse*, *H2ab*, *H2eb* and *H2dma* (Fig. 2b). Genes involved in detecting and transducing pro-inflammatory cues—such as *Tlr7*, *Tlr12*, *Nlrc5*, *Stat2*, *Stat6*, *Irf1* and *Irf7*—were also downregulated, as were several genes downstream of interferon signalling. Among the genes induced by isoDCA treatment, we identified negative regulators of NFκB, mitogen-activated protein kinase (MAPK) and cytokine-receptor signalling, including *Nfkbia*, *Dusp1*, *Dusp5* and *Socs1*. Given that the transcriptional profile imposed by isoDCA suggested an overall anti-inflammatory state, we tested the effects of this bile acid on the ability of DCs to prime antigen-specific T cells and to secrete cytokines in response to microbial cues. IsoDCA treatment decreased the production of the inflammatory cytokines tumour necrosis factor (TNF)-α and interleukin (IL)-6 upon agonist stimulation of Toll-like receptors

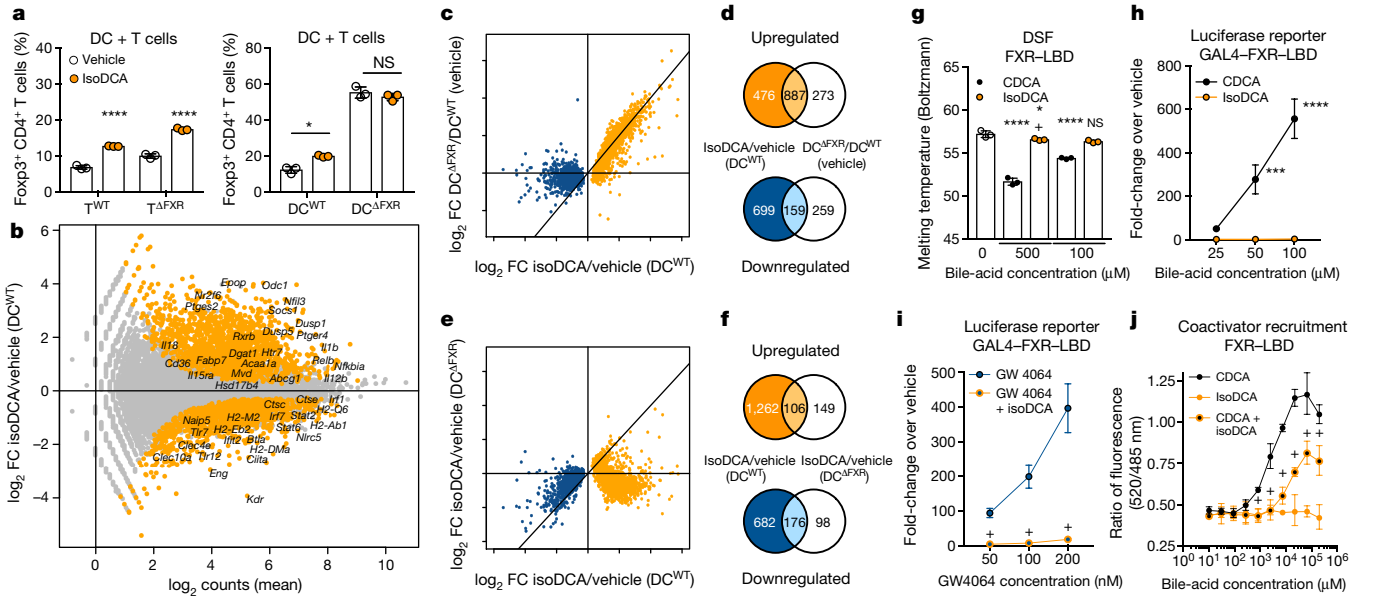


Fig. 2 | Potentiation of T_{reg} cell generation by isoDCA requires FXR expression in DCs. a, Effects of FXR deficiency in T cells (wild-type, WT, and FXR-deficient, ΔFXR) (left) or DCs (right) on T_{reg} cell induction by isoDCA (50 μM). **b**–**f**, RNA-seq analysis of FACS-purified DCs 24 h after exposure to bile acids. **b**, Transcriptional profiling of WT DCs treated with isoDCA (50 μM). Differentially expressed genes (adjusted *P*-value < 0.05) are in orange. **c**, Fold-change (FC) versus FC plot comparing the transcriptional changes induced by isoDCA treatment (*x*-axis) and FXR deficiency (*y*-axis). Genes downregulated by isoDCA are in blue; genes induced by isoDCA are in orange. **d**, Overlap between genes regulated by isoDCA treatment (orange) and FXR deficiency (blue). **e**, FC versus FC plot comparing the effects of isoDCA on WT (*x*-axis) and FXR-deficient (*y*-axis) DCs, colour-coded as in **c**. **f**, Overlap between genes regulated by isoDCA in WT (orange) and FXR-deficient (blue) DCs. **g**, Differential scanning fluorimetry (DSF) experiment with recombinant FXR–LBD and bile acids at 1,000-fold (500 μM) or 200-fold (100 μM) excess. Shown is the Boltzmann melting temperature of the FXR–LBD. **h**, **i**, Luciferase reporter assays. Cells expressing a Gal4–FXR–LBD fusion protein were treated with vehicle, CDCA, or GW4064 alone or combined with isoDCA (100 μM). **j**, FRET-based coactivator-recruitment assay. The indicated bile acids were mixed with glutathione-*S*-transferase (GST)-tagged FXR–LBD (5 nM), fluorescein isothiocyanate (FITC)–SRC2-2 coactivator peptide (500 nM) and GST antibody (5 nM). Shown is the ratio of fluorescence at 520 nm/485 nm. **a**, **g**, **j**, Displaying means ± s.d. of technical replicates (**a**, **g**, $n = 3$; **j**, $n = 4$), representative of three independent experiments. Data in **b**–**f** ($n = 3$) are from one experiment. Data in **h**, **i** ($n = 3$) are shown as means ± s.d., pooled from three independent experiments. Statistical significance determined by a one-way (**g**) or two-way (**a**, **h**–**j**) ANOVA followed by Tukey's or Sidak's test. $^{*}P < 0.05$, $^{***}P < 0.001$, $^{****}P < 0.0001$ vs vehicle; plus symbol shows $P < 0.05$ versus paired concentration of agonist; ns, not significant.

(TLRs) (Extended Data Fig. 3a). We further used a reporter T-cell line, in which green fluorescent protein (GFP) is under the control of NFAT (the nuclear factor of activated T cells), and which expresses OT-II—an MHC-II-restricted, ovalbumin-specific T-cell antigen receptor (TCR). We found that isoDCA reduced the DC-mediated activation of this cell line after pulse treatment with ovalbumin (Extended Data Fig. 3b), indicating broad anti-inflammatory activities of this bile acid.

Given our results implicating FXR in induction of T_{reg} cells by isoDCA, we compared the transcriptional changes elicited by FXR deficiency and exposure to bile acid. Genes that were significantly modulated in response to isoDCA treatment were, by and large, differentially expressed between FXR-deficient and -sufficient cells (Fig. 2c), indicating that these two perturbations lead to similar transcriptional changes in DCs. Notably, more than 54% of genes induced by isoDCA treatment were also significantly upregulated in FXR-deficient DCs compared with wild-type cells (Fig. 2c, d), implying that isoDCA may counteract FXR-mediated transcriptional repression. Although transcripts downregulated in response to isoDCA were generally present at lower levels in FXR-deficient DCs relative to wild-type cells (Fig. 2c), their expression was further decreased when FXR-deficient DCs were treated with isoDCA (Fig. 2e, f). Thus, while FXR may contribute to driving the expression of genes downregulated by bile acid treatment, these results suggest that additional, FXR-independent mechanisms also regulate these targets. We next sought to characterize the molecular interaction between FXR and isoDCA. While the naturally occurring FXR agonist chenodeoxycholic acid (CDCA)[22] induced a substantial shift in the melting temperature of recombinant FXR ligand-binding domain (LBD), isoDCA produced a less pronounced change, only evident at a high ligand-to-protein ratio (Fig. 2g). In line with this observation, CDCA elicited a robust signal in an FXR luciferase reporter assay, but isoDCA treatment had no effect (Fig. 2h). These results suggest a distinct mode of interaction between FXR and isoDCA, and raise the possibility of functional antagonism of this nuclear receptor. Indeed, we observed that isoDCA reduced the luciferase reporter signal in response to a synthetic FXR agonist, GW4064 (Fig. 2i). Corroborating these data, we found that isoDCA limited CDCA-induced FXR activity in a fluorescence resonance energy transfer (FRET)-based coactivator-recruitment assay (Fig. 2j). Together, these findings suggest that antagonizing the FXR-dependent transcriptional output of APCs may contribute to the pT_{reg} cell-inducing effects of isoDCA.

We then set out to explore the biological effects of isoDCA in vivo. Production of isoDCA from cholic acid involves chemical transformations performed by at least two different bacteria. The capacity for cleavage of the 7α-hydroxyl group from cholic acid has been observed in *Clostridium scindens*[8], while epimerization of the 3α-hydroxyl group of DCA was characterized in *Ruminococcus gnavus*[11] (Fig. 3a). To assess the effects of bacterial transformation of bile acids in the colon, we reconstructed the isomerization pathway for isoDCA by inserting hydroxysteroid dehydrogenases from *R. gnavus* (*Rumgna_02133* and *Rumgna_00694*) into *Bacteroides thetaiotaomicron* (*B. theta*), a genetically tractable commensal (*B. theta*^{eWT}) (Fig. 3b). The active site of the enzyme encoded by *Rumgna_00694* contains a tyrosine residue that is predicted to be conserved by homology modelling, which we changed to phenylalanine to create a catalytically dead mutant (*B. theta*^{eCD})

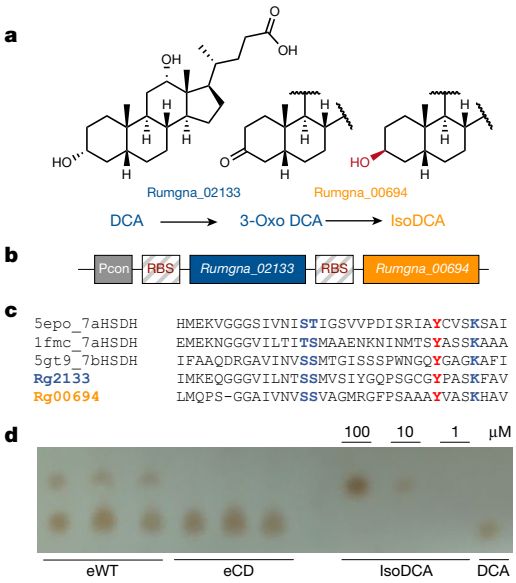


Fig. 3 | Engineering an isoDCA-producing strain of *Bacteroides thetaiotaomicron* (*B. theta*). a, Enzymes involved in isoDCA formation from DCA. Rumgna_02133 and Rumgna_00694, two hydroxysteroid dehydrogenases (HSDHs) present in *Ruminococcus gnavus*, were identified in ref. [11] as key enzymes catalysing epimerization of the 3-OH group of DCA by this bacterium. **b**, Cloning strategy to reconstitute the pathway for isoDCA generation in *B. theta*. Constructs for *Rumgna_02133* and *Rumgna_00694* were codon-optimized, put under the control of a strong, constitutive promoter (Pcon) in *B. theta* and chromosomally integrated by conjugation. RBS, ribosome-binding site. **c**, Rationale for the design of a catalytically dead (eCD) mutant of Rumgna_00694 (Rg00694). Alignment of partial amino-acid sequences of bacterial HSDHs: 5epo_7aHSDH from *Clostridium absonum*; 1fmc_7aHSDH from *Escherichia coli*; 5gt9_7bHSDH from *Collinsella aerofaciens*. A tyrosine (Y, red) found between additional conserved residues (blue) in the putative active site of HSDHs was mutated to phenylalanine (Y165F). **d**, Characterization of the biochemical activity of *B. theta*^{eWT} and *B. theta*^{eCD} strains by thin-layer chromatography (TLC). Bacteria were grown to exponential phase and transferred to media containing DCA. Following incubation for 24 h, media was extracted in ethyl acetate and analysed by TLC. DCA and isoDCA controls are shown on the four rightmost lanes. *B. theta*^{eWT} (three leftmost lanes) converts DCA into isoDCA, while *B. theta*^{eCD} mutant (three middle lanes) shows no activity. Data are representative of two experiments.

(Fig. 3c). As expected, we observed robust production of isoDCA by the engineered functional *B. theta* strain (as assessed by thin-layer chromatography), while conversion of DCA by the corresponding mutant strain was undetectable (Fig. 3d and Extended Data Fig. 4a).

Bacteroides species lack 7α-dehydroxylation activity; therefore, we assembled an isoDCA-producing consortium by combining *C. scindens* with our engineered strains (Fig. 4a, b). Both wild-type and catalytically dead consortia colonized mice to similar levels (Extended Data Fig. 5). All colonization conditions increased the percentage of colonic T_{reg} cells, including the RORγt$^+$ cell subset (Fig. 4c and Extended Data Fig. 6a, b). Despite having similar frequencies of bulk Foxp3$^+$ cells, recipients of the functional consortium showed a significant increase in the RORγt$^+$ pT_{reg} cell subset compared with mice colonized with the catalytically dead ensemble (Fig. 4d). In agreement with the notion that *Clostridium* species and secondary bile acids are indigenous to the colon, we failed to detect differential levels of RORγt-expressing pT_{reg} cells in the mesenteric lymph node or small intestine lamina propria (Extended Data Fig. 6c, d). Although Foxp3$^-$ RORγt$^+$ CD4$^+$ T cells were robustly induced in the colon of mice conventionalized by faecal microbiota transplantation (FMT), this effector T cell population was comparable between recipients of the functional or catalytically dead consortia both at 10 days and at 4 weeks after colonization (Extended Data Fig. 6e, f), consistent with our in vitro finding that isoDCA had no substantial effect on the generation of T_H17 cells (Extended Data Fig. 1a). By contrast, the differences in frequencies of colonic RORγt$^+$ pT_{reg} cells persisted at this later time point, although they were less pronounced (Extended Data Fig. 6g).

To exclude potential effects of bacterial strain background and to generalize these initial findings, we engineered two additional species of intestinal commensals to produce isoDCA. For this purpose, we generated functional and catalytically dead strains of *Bacteroides fragilis* (*B. frag*^{eWT} and *B. frag*^{eCD} respectively) and *Bacteroides ovatus* (*B. ova*^{eWT} and *B.ova*^{eCD}) using the above strategy. All functional engineered *Bacteroides* strains converted DCA into isoDCA in vitro, with *B. frag*^{eWT} and *B. ova*^{eWT} producing more isoDCA compared with *B.theta*^{eWT} and yielding similar levels to *R. gnavus* as measured by liquid chromatography–mass spectrometry (Extended Data Fig. 7a). No isoDCA was detected in the culture supernatants of catalytically dead strains. Functional engineered *B. frag* and *B. ova* strains also induced higher frequencies of colonic RORγt$^+$ T_{reg} cells by comparison with their respective catalytically dead counterparts (Fig. 4e, f and Extended Data Fig. 7b), suggesting that this effect does not depend on a particular strain background and is unlikely to be caused by a potential 'feedback' of isoDCA on the bacterium. Notably, isoDCA levels in the caecal contents of animals colonized with a high-producing strain (*B. frag*^{eWT}) did not surpass that of mice receiving FMT (Extended Data Fig. 7c), indicating that our reconstruction of this enzymatic pathway did not lead to supraphysiological bile acid levels. Production of short-chain fatty acids (SCFAs) was comparable between consortia (Extended Data Fig. 8), suggesting that bile acid transformation did not broadly affect bacterial metabolism, and that isoDCA-producing bacteria were able to increase the number of pT_{reg} cells in the presence of other 'tolerogenic' metabolites.

Given that the hydroxysteroid dehydrogenases introduced into *Bacteroides* could potentially modify substrates other than DCA, we tested the effects of our engineered bacteria in the absence of the 7α-dehydroxylating commensal—that is, without *C. scindens*. Mono-colonization of germ-free mice with functional or catalytically dead *B. frag* or *B. ova* resulted in similar frequencies of colonic RORγt$^+$ pT_{reg} cells (Fig. 4g,h), demonstrating that the biological activity of our isoDCA-producing consortia depends on the presence of a DCA-generating bacterium. To confirm that our engineered strains promote bona fide pT_{reg} cell generation in vivo, we colonized germ-free CNS1-sufficient (*Foxp3*^{GFP}) and -deficient (*Foxp3*$^{GFP\Delta CNS1}$) mice with functional or catalytically dead consortia. By comparison with catalytically dead controls, colonization with functional consortia led to much higher frequencies of Foxp3$^+$ RORγt$^+$ pT_{reg} cells in CNS1-sufficient animals (Fig. 4i, j). CNS1-deficient hosts with genetically impeded pT_{reg} cell generation showed similarly low frequencies of Foxp3$^+$ RORγt$^+$ CD4$^+$ T cells when colonized with functional or catalytically dead consortia (Fig. 4i, j). These experiments show that colonization with isoDCA-producing microbial consortia promotes de novo generation of colonic pT_{reg} cells.

Previously, pT_{reg} cells were shown to dampen immune responses during microbial colonization and to support the metabolic function of the gut microbiota[23]. The establishment of such immunological tolerance to commensals along with other types of host–microbe interaction probably evolved around conserved features of microbial communities, such as their metabolic output. Supporting this notion, we have found that, in addition to bacterial fermentation products[5], the secondary bile acid isoDCA is also a potent inducer of pT_{reg} cells. Using engineered *Bacteroides* strains as part of rationally designed, minimal microbial consortia, we have shown that isoDCA-producing bacteria promote the generation of pT_{reg} cells in vivo in a CNS1-dependent manner. IsoDCA limited FXR activity in DCs and conferred upon them an anti-inflammatory phenotype. Although our data support the involvement of myeloid-cell-intrinsic FXR activity in the induction of pT_{reg} cells,

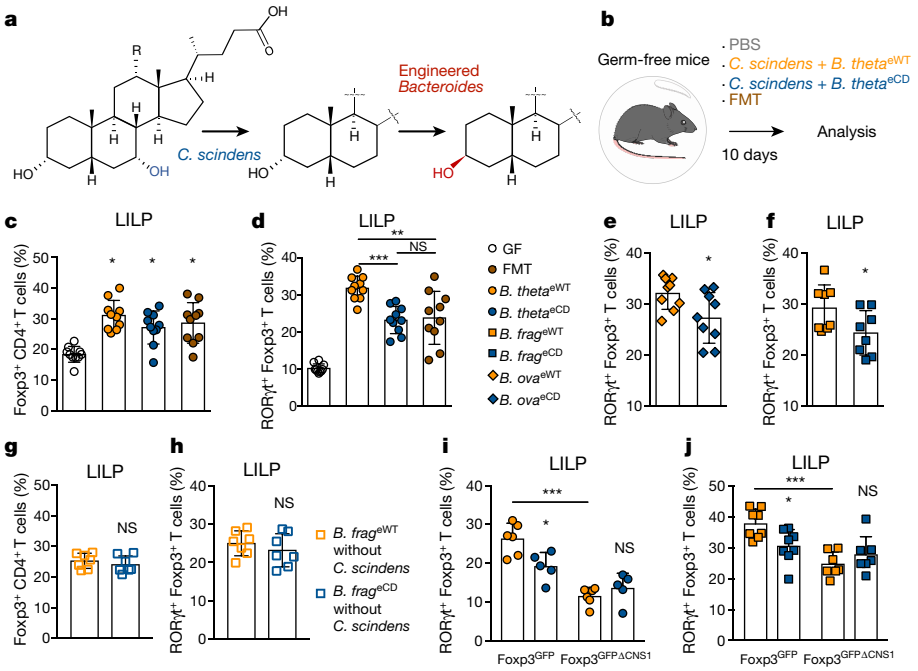


Fig. 4 | Defined bacterial consortia containing isoDCA-producing strains promote generation of pT$_{reg}$ cells in vivo. a, Generation of isoDCA by a minimal microbial consortium. Enzymatic steps performed by *C. scindens* and the engineered *Bacteroides sp.* (*B. thetaiotamicron*, *B. fragilis* and *B. ovatus*). **b**, Experimental setup. Germ-free mice were colonized with consortia containing *C. scindens* plus *B. theta*^{eWT} or *C. scindens* plus *B. theta*^{eCD}. Recipients of an FMT and non-colonized mice (treated with phosphate-buffered saline (PBS)) served as references. The immune-cell composition in the LILP was analysed on day 10 by FACS. **c**, **d**, Frequency of Foxp3$^+$ (**c**) and RORγt$^+$ (**d**) T$_{reg}$ cells. **e**, **f**, Frequency of RORγt$^+$ T$_{reg}$ cells in germ-free mice colonized with *C. scindens* in combination with engineered *B. ova* (**e**) or *B. frag* (**f**). **g**, **h**, *B. frag*^{eWT} or *B. frag*^{eCD} were administered to germ-free mice the absence of a 7-α-dehydroxylating bacterium (without *C. scindens*). Shown are frequencies of Foxp3$^+$ (**g**) and RORγt$^+$ (**h**) T$_{reg}$ cells. **i**, **j**, Germ-free *Foxp3*^{GFP} and *Foxp3*^{GFPΔCNS1} littermate mice were gavaged with *C. scindens* plus *B. theta*^{eWT} (**i**) or *C. scindens* plus *B. frag*^{eWT} (**j**) and compared with recipients of the respective CD consortia. Shown are percentages of colonic RORγt$^+$ Foxp3$^+$ T$_{reg}$ cells. **c**–**j**, Data combined from two independent experiments (except in **i**) and shown as means ± s.d. Statistical significance determined by two-tailed *t*-test (**e**, $n = 8$; f, $n = 9$; **g**, **h**, $n = 7$), or one-way (**c**, **d**, $n = 10$) or two-way (**i**, $n = 5$; j, $n = 7$) ANOVA followed by Tukey's multiple comparison's test. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; ns, not significant.

the relative contribution of this and other bile acid-sensing receptors in mediating the effects of isoDCA on the mucosal immune milieu remains to be investigated. In conclusion, our findings suggest that microbial metabolism of endogenous steroids contributes to immunological balance in the colon.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2193-0.

1. Josefowicz, S. Z. et al. Extrathymically generated regulatory T cells control mucosal TH2 inflammation. *Nature* **482**, 395–399 (2012).
2. Zheng, Y. et al. Role of conserved non-coding DNA elements in the Foxp3 gene in regulatory T-cell fate. *Nature* **463**, 808–812 (2010).
3. Nutsch, K. et al. Rapid and efficient generation of regulatory T cells to commensal antigens in the periphery. *Cell Rep.* **17**, 206–220 (2016).
4. Kim, K. S. et al. Dietary antigens limit mucosal immunity by inducing regulatory T cells in the small intestine. *Science* **351**, 858–863 (2016).
5. Arpaia, N. et al. Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature* **504**, 451–455 (2013).
6. Furusawa, Y. et al. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446–450 (2013).
7. Smith, P. M. et al. The microbial metabolites, short-chain fatty acids, regulate colonic T$_{reg}$ cell homeostasis. *Science* **341**, 569–573 (2013).
8. Ridlon, J. M., Kang, D. J. & Hylemon, P. B. Bile salt biotransformations by human intestinal bacteria. *J. Lipid Res.* **47**, 241–259 (2006).
9. Ridlon, J. M., Harris, S. C., Bhowmik, S., Kang, D. J. & Hylemon, P. B. Consequences of bile salt biotransformations by intestinal bacteria. *Gut Microbes* **7**, 22–39 (2016).
10. Wahlström, A., Sayin, S. I., Marschall, H. U. & Bäckhed, F. Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. *Cell Metab.* **24**, 41–50 (2016).
11. Devlin, A. S. & Fischbach, M. A. A biosynthetic pathway for a prominent class of microbiota-derived bile acids. *Nat. Chem. Biol.* **11**, 685–690 (2015).
12. Eyssen, H., De Pauw, G., Stragier, J. & Verhulst, A. Cooperative formation of omega-muricholic acid by intestinal microorganisms. *Appl. Environ. Microbiol.* **45**, 141–147 (1983).
13. Eyssen, H. J., De Pauw, G. & Van Eldere, J. Formation of hyodeoxycholic acid from muricholic acid and hyocholic acid by an unidentified gram-positive rod termed HDCA-1 isolated from rat intestinal microflora. *Appl. Environ. Microbiol.* **65**, 3158–3163 (1999).
14. Hamilton, J. P. et al. Human cecal bile acids: concentration and spectrum. *Am. J. Physiol. Gastrointest. Liver Physiol.* **293**, G256–G263 (2007).
15. Hofmann, A. F. & Small, D. M. Detergent properties of bile salts: correlation with physiological function. *Annu. Rev. Med.* **18**, 333–376 (1967).
16. Kretschmer, K. et al. Inducing and expanding regulatory T cell populations by foreign antigen. *Nat. Immunol.* **6**, 1219–1227 (2005).
17. Haxhinasto, S., Mathis, D. & Benoist, C. The AKT-mTOR axis regulates de novo differentiation of CD4+Foxp3+ cells. *J. Exp. Med.* **205**, 565–574 (2008).
18. Hang, S. et al. Bile acid metabolites control T$_H$17 and T$_{reg}$ cell differentiation. *Nature* **576**, 143–148 (2019).
19. Song, X. et al. Microbial bile acid metabolites modulate gut RORγ$^+$ regulatory T cell homeostasis. *Nature* **577**, 410–415 (2020).
20. Sefik, E. et al. Individual intestinal symbionts induce a distinct population of RORγ$^+$ regulatory T cells. *Science* **349**, 993–997 (2015).
21. Ohnmacht, C. et al. The microbiota regulates type 2 immunity through RORγt$^+$ T cells. *Science* **349**, 989–993 (2015).
22. Parks, D. J. et al. Bile acids: natural ligands for an orphan nuclear receptor. *Science* **284**, 1365–1368 (1999).
23. Campbell, C. et al. Extrathymically generated regulatory T cells establish a niche for intestinal border-dwelling bacteria and affect physiologic metabolite balance. *Immunity* **48**, 1245–1257 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Dendritic-cell isolation

B16 melanoma cells secreting Flt3 ligand (provided by G. Dranoff) were injected subcutaneously into the left flank of mice to expand splenic DCs in vivo. Ten to twenty days after tumour injection, spleens were harvested and dissociated in RPMI 1640 medium containing 1.67 U ml^{-1} liberase TL (Roche) and 50 μg ml^{-1} DNase I (Roche) for 20 min at 37 °C with vigorous shaking (250 r.p.m.). Digested spleens were passed through a 100-μm strainer and washed in complete RPMI (RPMI 1640 with 10 mM HEPES buffer (ThermoFisher), 1% penicillin/streptomycin (ThermoFisher), 1% L-glutamine (ThermoFisher) and 10% fetal bovine serum (FBS, ThermoFisher). Single-cell suspensions were enriched for DCs using a MACS CD11c Microbeads Ultrapure isolation kit (Miltenyi Biotec) according to the manufacturer's instructions. Purity of DC enrichment was assessed by FACS analysis (typical purity: greater than 92% $CD11c^+$ $MHCII^{hi}$ cells).

In vitro assays

Polarization of T cells. Naive ($CD44^-$ $CD62L^+$ GFP^-) $CD4^+$ T cells from *Foxp3*^{GFP} mice were FACS-purified from spleen and peripheral (pooled inguinal, brachial, axial and submandibular) lymph nodes after a CD4-enrichment step (Dynabeads 11461D, Invitrogen), performed as per the manufacturer's instructions. DC (1×10^5) plus naive $CD4^+$ T cell (5×10^4) cocultures were set up in the presence of 1 μg ml^{-1} of monoclonal CD3ε antibody (InvivoMAb, BioXcel). Cholic acid (catalogue number C19000-000), chenodeoxycholic acid (C0985-000), lithocholic acid (C1420-000), isolithocholic acid (C10475-00), deoxycholic acid (C1070-000), isodeoxycholic acid (C1165-000), 3-oxodeoxycholic acid (C1725-000), ursodeoxycholic (C1020-000), α-muricholic acid (C1890-00), β-muricholic acid (C1895-000), γ-muricholic acid (C1850-000) and ω-muricholic acid (C1888-000) were all purchased from Steraloids Inc. (Newport, RI). 6-Oxo-muricholic acid (5β-cholanic acid-3α,7β-diol-6-one) was produced by the Organic Chemistry Synthesis core at the Memorial Sloan Kettering Cancer Center (MSKCC). We stored 100 mM stocks of bile acids in dimethylsulfoxide (DMSO) at −80 °C. For T-cell activation in the absence of DCs, naive $CD4^+$ T cells (5×10^4) were incubated with mouse T-activator CD3/CD28 Dynabeads (Gibco) at a 1-to-1 bead-to-cell ratio. For assessment of cell proliferation, naive $CD4^+$ T cells were labelled with 5 μM Cell Trace Violet (Invitrogen) according to the manufacturer's instructions. Suboptimal T_{reg}-induction conditions consisted of 1 ng ml^{-1} recombinant human TGF-β1 (R&D) and 100 U ml^{-1} recombinant human IL-2 (Biological Resources Branch, NCI). For T_H17 differentiation, cells were incubated with 2 ng ml^{-1} recombinant human TGF-β1 (R&D) and 20 ng ml^{-1} recombinant murine IL-6 (PeproTech). All in vitro polarization assays were carried in complete RPMI with 10% FBS (final volume of 200 μl) in flat-bottom 96-well plates (USA Scientific). Border wells were filled with media only (no cells) to minimize evaporation during incubation at 37 °C/5% CO_2. On day 3 of culture, cells were transferred into V-bottom plates (Fisher), pelleted by centrifugation and incubated with antibody staining mix containing Ghost Dye Red 780 viability dye diluted in PBS for 15 min at 4 °C. For cytokine production analyses, cells were incubated for 3 h at 37 °C/5% CO_2 in restimulation media (complete RPMI 1640 with 5% FBS, 50 ng ml^{-1} PMA (Sigma), 500 ng ml^{-1} ionomycin (Sigma), 1 μg ml^{-1} brefeldin A (Sigma) and 2 μM monensin (Sigma)). Extracellular antigens were stained for 15 min at 4 °C with an antibody staining mix containing Ghost Dye Red 780 viability dye diluted in PBS. Cells were fixed and permeabilized with BD Cytofix/Cytoperm for 20 min at 4 °C. Antibodies against intracellular antigens were diluted in 1× BD Perm/Wash buffer and cells were stained for 30 min at 4 °C. Cytometry data were acquired on a LSRII digital cell analyser (Becton Dickinson, NJ). Foxp3 induction was assessed by expression of GFP reporter protein. 123count eBeads (Invitrogen) were added at 5,000 beads per sample to quantify absolute cell numbers.

TLR agonist activation. DCs (1×10^5 per well, flat-bottom 96-well plates) were pretreated with isoDCA or vehicle in complete RPMI 10% FBS for 5 h before stimulation with various TLR ligands (Pam3CSK4, 0.5 μg ml^{-11}; HKLM, 10^7 cells ml^{-1}; poly(I:C) and poly(I:C) LMW, 5 μg ml^{-1}; LPS-EK, 5 μg ml^{-1}; ST-FLA, 1 μg ml^{-1}; and ODN1826, 2.5 μM) from the mouse TLR1–9 agonist kit (Invivogen tlrl-kit1mw) for 18 h at 37 °C/5% CO_2. TNF-α and IL-6 levels were quantified in cell-free supernatants (roughly 150 μl of 200 μl cell culture after centrifugation) by enzyme-linked immunosorbent assay (ELISA; eBioscience, catalogue numbers 5017331 and 5017218) according to the manufacturer's instructions.

Antigen-processing and -presentation assays. DCs (1×10^4) were pulsed with ovalbumin (Sigma A5503, 1 mg ml^{-1}) or CD3ε antibody (1 μg ml^{-1}) in the presence or absence of isoDCA for 1 h at 37 °C/5% CO_2 in serum-free complete RPMI. Then, an equal volume of 10% FBS RPMI was added to a final serum concentration of 5%. After 5 h, we added 1×10^4 reporter T cells (TCR-α/β-null, BW5147 mouse thymoma cells[24] with NFAT-controlled GFP expression[25]) carrying the OTII TCR, along with ovalbumin or CD3ε antibody and isoDCA (to keep concentrations constant) in complete RPMI with 10% FBS. The frequency of GFP^+ cells was determined 24 h later by FACS. Reporter cells were not tested for mycoplasma or subjected to additional validation other than functional assessment during experiments.

RNA-sequencing of dendritic cells

In vitro DC/T-cell coculture assays in the presence of bile acids were carried out as above. Following 24 h of incubation at 37 °C in 5% CO_2, cells were transferred into V-bottom plates, pelleted by centrifugation and incubated with antibody staining mix containing Ghost Dye Red 780 viability dye diluted in PBS for 15 min at 4 °C. Approximately 3×10^4 DCs were double-sorted on a FACSAria II (Becton Dickinson, NJ) on the basis of viability and CD11c/MHC II expression. Purified cells were resuspended in Trizol and submitted for RNA-sequencing at the Integrated Genomics Core (iGO) at MSKCC. Samples underwent SMARTer amplification (Takara) and were sequenced on a Hiseq platform (Illumina) at a depth of 20 million to 30 million paired-end 50-base-pair reads per sample. RNA-sequencing reads were aligned to the reference mouse genome (Gencode m19) using STAR RNA-Seq aligner[26], and read counts were obtained with HTSeq-count[27]. The DESeq2 R package[28] was used to perform differential gene expression analyses between groups. A cutoff of 0.05 was set on obtained *P*-values (adjusted using the Benjamini–Hochberg correction for multiple comparisons) to define statistically significant genes for each comparison.

Luciferase assays

Reporter cells expressing GAL4–LBD fusions for human FXR (Indigo Biosciences, IB006001) were incubated for 22–24 h in the presence of CDCA, the synthetic agonist GW4064 (provided with kit) and/or isoDCA. Cells were processed according to the manufacturer's instructions, including the recommended step to determine viability using the Live Cell Multiplex assay (LCM, Indigo Biosciences). Luminescence was read in a GloMax 96 Luminometer (Promega).

Differential scanning fluorimetry assay

IsoDCA and CDCA were tested for their ability to alter the melting temperature of recombinant FXR–LBD (Invitrogen) using a modified version of the protocol in ref. [29]. Briefly, bile acids and FXR–LBD (final concentration 500 nM) were diluted in assay buffer (10 mM Tris (pH 8.3), 0.5 mM EDTA, 100 mM NaCl and 5 mM DTT, added fresh before assay from a ×100 frozen stock) and combined into transparent 384-well plates (Applied Biosystems). SYPRO Orange dye (Protein Thermal Shift

Dye, Applied Biosystems) was added at a final concentration of 1/2,000 in a reaction volume of 20 µl. Temperature-dependent changes in fluorescence were detected in a QuantStudio 6 Flex real-time quantitative polymerase chain reaction (qPCR) system (Applied Biosystems). The equipment was programmed with the following thermal profile: step 1, temperature 25 °C, time 2 min; step 2, temperature 99 °C, time 2 min; continuous ramp mode; ramp rate for step 1, 1.6 °C s^{-1}; ramp rate for step 2, 0.05 °C s^{-1}. Data analysis was performed using PTS software (Applied Biosystems).

FRET-based FXR coactivator recruitment assay

IsoDCA and CDCA were tested for their ability to activate FXR in a cell-free FRET assay using LanthaScreen technology according to the manufacturer's protocol (Thermo Fisher, PV4833). Briefly, after combining diluted bile acids with GST-tagged FXR–LBD, terbium GST antibody and fluorescently labelled SRC2 in white, flat-bottom 384-well plates (Greiner Bio-one), the reaction was incubated at room temperature in the dark under gentle shaking (60 r.p.m.) for 1 h before reading. Fluorescence detection was carried out in a Tecan Infinite M100 Pro Microplate reader (Tecan Group, Switzerland) set up according to the LanthaScreen Terbium Assay Setup guide available at www.lifetechnologies.com/instrumentsetup. For the first fluorescence reading, settings were as follows. Wavelength: excitation 332 nm, bandwidth 20.0 nm; emission 485 nm; bandwidth 20.0 nm. Flashes: mode 2 (100 Hz (20)), settle time 0 ms. Mode: top. Gain: optimal. *Z*-position: calculated from well (select well with appropriate substrate). Integration: lag time 100 µs, integration time 200 µs. After a second fluorescence reading was added to the existing protocol, settings were adjusted to those described above, except for the following. Wavelength: excitation 332 nm, bandwidth 20.0 nm; emission 515 nm; bandwidth 20.0 nm. Results were expressed as ratios of fluorescence at 520 nm to 485 nm.

16S amplicon sequencing of intestinal microbiota

Sample prep, sequencing and OTU clustering were performed by the Microbiome Core Lab at Weill Cornell Medicine. DNA extraction from caecal contents was carried out with a Maxwell RSC PureFood GMO and authentication kit (Promega, AS1600). Preweighted faecal material was deposited into a PowerBead glass 0.1-mm tube (Qiagen, 13118-50) and 1 ml of CTAB buffer plus 20 µl of RNase A solution were added. Samples were homogenized in buffer by gentle vortexing for 10 s and then incubated for 5 min in a ThermoMixer F2.0 (Eppendorf), shaking at 1,500 r.p.m. Homogenized samples were then vortexed horizontally at high -speed for 10 min. Samples were centrifuged at 12,700 r.p.m. in an Eppendorf centrifuge for 10 min and extraction proceeded in an automated platform. Tubes were transferred to a MaxPrep Liquid Handler tube rack (Promega) and the Maxwell RSC48 instrument (Promega, AS8500) was loaded with proteinase K tubes, lysis buffer, elution buffer, pipetting tips, 96-sample deep-well plate and Maxwell RSC 48 sheath tips. The instrument was programmed to use 300 µl of sample and transfer all sample lysate into the Maxwell RSC 48 extraction cartridge for DNA extraction. Upon completion, the extraction cartridge was loaded into Maxwell RSC 48 for DNA extraction and elution. DNA was eluted in 100 µl and transferred into 96-well plates. A Quant-iT double-stranded (ds)DNA high-sensitivity assay (Thermo Fisher, Q33120) was used for DNA quantification. 16S libraries were generated according to the Earth Microbiome Project available at http://press.igsb.anl.gov/earthmicrobiome/protocols-and-standards/16s/. Briefly, the V4 region was amplified using a Hot Start PCR mix (ThermoFisher, 13000014). Primers were modified from ref. [30] (barcodes were moved to the 515F primer) are as follows: 515F forward primer, barcoded (XXXXXXXXXXXX) AATGATAC GGCGACCACCGAGATCTACACGCT XXXXXXXXXXXX TATGGTAATT GT GTGYCAGCMGCCGCGGTAA; 806R reverse CAAGCAGAAGACGG CATACGAGAT AGTCAGCCAG CC. Cycling conditions for 96-well-plate thermocyclers were: 3 min at 94 °C; (45 s at 94 °C; 60 s at 50 °C; 90 s at 72 °C) ×35; 10 min at 72 °C; hold at 4 °C.

PCR products were purified with a MoBio UltraClean PCR clean-up kit (catalogue number 12500) according to the manufacturer's instructions. Library concentrations were determined with a Quant-iT dsDNA high-sensitivity assay. Verification of library quality and size was performed using a PerkinElmer LabChip GXII instrument with DNA 1K reagent kit (CLS760673). Libraries were normalized and pooled at 2 nM before sequencing on a MiSeq instrument (Illumina) at a loading concentration of 5.5 pM with 10% PhiX, paired-end 250 using MiSeq reagent kit v2, 500 cycles (MS-102-2003). Custom sequencing primers were as follows: read 1, TATGGTAATT GT GTGYCAGCMGCCGCGGTAA; index, AATGATACGGCGACCACCGAGATCTACACGCT; read 2, AGTCAGCCAG CC GGACTACNVGGGTWTCTAAT. Note that the 5′-adaptor sequence/index sequencing primer has an extra GCT at its 3′-end compared with Illumina's usual index primer sequences. These bases were added to the 3′-end of the Illumina 5′-adaptor sequence to increase the melting temperature for read 1 during sequencing.

For analysis, demultiplexed raw reads were processed to generate an operational taxonomic unit (OTU) table using USEARCH version 11.0.667 (ref. [31]). Specifically, forward and reverse reads were merged using a maximum of five mismatches in the overlap region, a minimum sequence identity in the overlap region of 90%, a minimum overlap length of 16 base pairs, and a minimum merged sequence length of 300 base pairs. PhiX contamination was then removed, followed by quality filtering based on FASTQ quality scores, with a maximum expected error number of 1.0. OTU clustering was performed using *usearch -cluster_otus* with default settings. Merged (prefilter) reads were mapped to the OTU sequences to generate the OTU table. Taxonomic classification of OTU representative sequences was performed using *usearch-sintax*, an implementation of the SINTAX algorithm[32] using version 16 of the Ribosomal Database Project (RDP) training set[33].

Engineering recombinant *Bacteroides* strains

The *3α-HSDH Rg_2133* and *3β-HSDH Rg_00694* genes were synthesized (GeneWiz) after codon optimization for expression in *B. theta*. A Golden Gate assembly kit (NEB) was used to assemble the following fragments and insert them into plasmid pNBU2 (ref. [34]) in 5′ to 3′ order: ppWW3806, ribosome binding site 8 (RBS), Rg_2133 open reading frame, RBS 8, Rg_00694 and pWW3810. The assembled construct was transformed into Stellar competent cells (Takara) and positive clones were selected and verified by Sanger sequencing. The plasmid was purified and retransformed into *E. coli* S17-1 cells for conjugation into *B. theta*, *B. frag* and *B. ova*. Mid-log *Bacteroides* cultures and plasmid-containing *E. coli* were mixed 10/1, centrifuged and allowed to sit for 20 min before resuspending the pellet in PBS and spreading on brain heart infusion (BHI) agar with 10% horse blood containing 25 µg ml^{-1} erythromycin and 200 µg ml^{-1} gentamycin, followed by incubation in an anaerobic chamber (Coy Labs) for 48 h at 37 °C. Individual colonies were screened by Sanger sequencing. For more information on bacterial strains and plasmids, see Supplementary Table 1.

Thin-layer chromatography

Strains were grown for 24 h or 72 h in yeast casitone fatty acids broth with carbohydrate (YCFAC) liquid medium (Anaerobe Systems) containing 100 µM DCA. We added 2 g NaCl to 7 ml culture supernatant, followed by 700 µl 6M HCl. The supernatant was extracted with 5 ml ethyl acetate, dehydrated in 2 g $MgSO_4$ and then extracted again with ethyl acetate. The organic extract was filtered through 40-µm nylon filters (Falcon), then dried by vacuum centrifugation in glass tubes. The pellet was resuspended in methanol, tubes were washed in an equal volume of acetone and the solvent was again removed by vacuum centrifugation. The final dried pellet was resuspended in 100 µl acetone and was spotted onto glass-backed silica plates (Sigma), which were developed in 70/20/2 benzene/1,4-dioxane/acetic acid. Plates were stained in 10% *w*/*v* $CuSO_4$ 8% H_3PO_4 and dried over a hot plate.

Mass spectrometry

IsoDCA quantification. Bacterial broth was mixed with an equal volume of methanol and centrifuged at 21,000*g* for 20 min. For in vivo quantification, about 10–30 mg of the dried faecal pellet was extracted in 150 µl of 50% methanol in water and vortexed for 15 min. The mixture was then spun down at 21,000*g* for 15 min. The supernatant of bacterial broth or extracted faecal pellet was analysed using an Agilent 1290 LC system coupled to an Agilent 6530 quadripole time-of-flight (QTOF) mass spectrometer with a 2.1 µm, 2.1 × 50 mm Zorbax Eclipse Plus C18 column (Agilent). Water with 0.05% formic acid (A) and acetone with 0.05% formic acid (B) was used as the mobile phase at a flow rate of 0.40 ml min^{-1} over a 11-min gradient: 0–1 min, 0% B; 1–3 min, 0–40% B; 3–10 min, 40–100% B; 10–11 min, 100–0% B. Reagents were mass-spectrometry grade, purchased from Fisher Scientific. All data were collected in negative-ion mode.

SCFA quantification. Caecal samples (roughly 120 mg) were weighed into 2-ml microtubes containing 2.8-mm ceramic beads (Omni International) and resuspended to a final concentration of 100 mg ml^{-1} using 80/20 methanol/water containing acetate-d3, propionate-d5, butyrate-d7 and valerate-d9 as internal standards (Cambridge Isotope Laboratories). Homogenization was carried out using a Bead Ruptor (Omni International) at 5.4 m s^{-1} for 3 min at 4 °C. Samples were centrifuged for 20 min at 20,000*g* at 4 °C, and 100 µl of caecal extract was added to 100 µl of 100 mM borate buffer (pH 10). Subsequently, 400 µl of 100 mM pentafluorobenzyl bromide (Thermo Scientific) diluted in acetonitrile (Fisher) and 400 µl of cyclohexane (Acros Organics) were added and reaction vials were sealed. Samples were heated to 65 °C for 1 h with agitation and then cooled to room temperature and centrifuged to promote phase separation. We transferred 100 µl of the cyclohexane (upper) phase to a new autosampler vial and carried out analysis at 1/10 and 1/100 dilutions (using cyclohexane). A calibration curve and quality-control samples were prepared in borate buffer covering the range 0.1–50 mM. Analysis by gas chromatography-mass spectrometry (GC-MS) was using an Agilent 7890A gas chromatograph and Agilent 5975C MS detector operating in negative chemical ionization mode. Methane was used as the chemical-ionization reagent gas at 2 ml min^{-1}, and a 1 µl splitless injection was made onto a HP-5MS column (30 m × 0.25 mm, 0.25 µm; Agilent Technologies). For SCFA quantification, the raw peak areas of acetate (*m*/*z* 59) and propionate (*m*/*z* 73) were normalized to acetate-d3 (*m*/*z* 62) and propionate-d5 (*m*/*z* 78) internal standards respectively; the C4 compounds butyrate and isobutyrate (*m*/*z* 87) were normalized to butyrate-d7 (*m*/*z* 94) and the C5 compounds 2-methylbutyrate, valerate and isovalerate (*m*/*z* 101) were normalized to valerate-d9 (*m*/*z* 110). Data analysis was performed with Agilent MassHunter quantitative analysis software (version 10.1, Agilent Technologies).

Colonization with engineered consortia

Frozen stocks of *Clostridium scindens* (ATCC, catalogue number 35704) or engineered *Bacteroides* strains were streaked on Columbia agar plates (BD) and grown at 37 °C inside an anaerobic chamber. Bacteria were scraped from agar plates with bacteriological loops into sterile anaerobic PBS. Consortia were assembled and transported in airtight tubes to the animal facility and administered to experimental animals by oral gavage. Mice were housed in flexible PVC isolators (Park Bioservices) or sentry sealed positive pressure cages (SPP, Allentown) for the duration of the experiments. For experiments in SPP cages, animals were manipulated with sterile gloves using aseptic techniques inside biosafety cabinets.

Isolation of cells from intestinal lamina propria

Approximately 12 cm of the distal small intestine and the combined caecum and colon were processed as the small intestine and large intestine, respectively. After removal of adherent adipose tissue and resection of Peyer's patches, intestines were opened longitudinally and shaken vigorously in PBS to release contents. Tissues were incubated in 25 ml intestinal intraepithelial lymphocyte (IEL) solution (1× PBS with 2% FBS (ThermoFisher), 10 mM HEPES buffer (ThermoFisher), 1% penicillin/streptomycin (ThermoFisher), 1% L-glutamine (ThermoFisher), plus 1 mM EDTA (Sigma) and 1 mM dithiothreitol (DTT; Sigma) added immediately before use) for 15 min at 37 °C with vigorous shaking (250 r.p.m.). Intestines were removed from IEL suspension, rinsed in PBS and transferred into 50-ml tubes containing 3× one-quarter-inch ceramic beads (MP Biomedicals) and 25 ml collagenase solution (1× RPMI 1640 with 2% FBS (ThermoFisher), 10 mM HEPES buffer (ThermoFisher), 1% penicillin/streptomycin (ThermoFisher), 1% L-glutamine (ThermoFisher), 1 mg ml^{-1} collagenase A (Sigma) and 1 U ml^{-1} DNase I (Sigma)). Following incubation for 30 min at 37 °C with vigorous shaking (250 r.p.m.), digested lamina propria samples were passed through a 100-µm strainer and centrifuged to remove collagenase solution. The lamina propria fractions were washed by centrifugation (5 min at 450*g*) in 44% Percoll (ThermoFisher) in PBS to remove debris and excess epithelial cell contamination before ex vivo restimulation for analysis of cytokine production. Foxp3 expression was assessed by detection of the fluorescent GFP reporter protein or intracellular staining using a Foxp3/transcription factor staining buffer set (eBioscience). 123count eBeads (Invitrogen) were added at 5,000 beads per sample to determine cell numbers. Cytometry data were acquired on an LSRII (Becton Dickinson, NJ).

Mice

Germ-free C57Bl/6 mice were purchased from Taconic and maintained in semirigid isolators (Park Bioservices) at Boehringer Ingelheim (Ridgefield, CT). Female mice were used in all experiments to facilitate distribution of animals into experimental groups. Germ-free *Foxp3*$^{GFP\Delta CNS1}$ and *Foxp3*^{GFP} mice were rederived as described[23] and maintained in flexible isolators (Class Biologically Clean, Madison, WI) at Weill Cornell Medicine. Male and female littermate mice were used in all experiments. Animals were fed with autoclaved 5KA1 chow. Germ-free status was routinely checked by aerobic and anaerobic cultures of faecal samples for bacteria and fungi and by PCR of faecal DNA samples for bacterial 16S and fungal/yeast 18S genes. Germ-free mice were at least eight weeks old at the start of experiments.

Specific-pathogen-free mice were housed at the Research Animal Resource Center for MSKCC and Weill Cornell Medicine with 12-h light/dark cycles under ambient conditions and ad libitum access to food and water. Experimental mice were maintained in a standard rodent diet (5053, LabDiet). F. Gonzalez (NIH, USA) provided the *Nr1h4*^{fl} mouse strain; *Csf1r*^{cre} mice were provided by F. Geissman at MSKCC; *CD4*^{cre} mice (*Tg(Cd4-cre)1Cwi*) were purchased from Jax laboratories and maintained in house. Experimental littermate animals were generated by mating mice homozygous for the *Nr1h4*^{fl} allele, with one of the breeders (male or female) carrying one copy of the Cre-driver gene. Cells from male and female *CD4*^{cre} *Nr1h4*^{fl} or *Csf1r*^{cre} *Nr1h4*^{fl} mice were used for in vitro experiments. Male *Csf1r*^{cre} *Nr1h4*^{fl} mice were analysed at six to eight weeks of age.

All studies were carried out under protocol 08-10-023 and approved by the Sloan Kettering Institute Institutional Animal Care and Use Committee. Germ-free mice housed at Boehringer Ingelheim were maintained under a protocol approved by the Institutional Animal Care and Use Committee of Boehringer Ingelheim Pharmaceuticals. All animals used here had no previous history of experimentation and were naive at the time of analysis.

Statistical analyses

Statistical tests were performed with GraphPad Prism version 7.0.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

RNA-sequencing and 16S amplicon sequencing data are available under BioProject (https://www.ncbi.nlm.nih.gov/bioproject/) identification codes PRJNA600898 and PRJNA600979. Source data for Figs. 1–4 and Extended Data Figs. 1–8 are available as .xsl tables with the paper. Other relevant data are available from the corresponding authors upon reasonable request.

24. White, J. et al. Two better cell lines for making hybridomas expressing specific T cell receptors. *J. Immunol*. **143**, 1822–1825 (1989).
25. Hooijberg, E., Bakker, A. Q., Ruizendaal, J. J. & Spits, H. NFAT-controlled expression of GFP permits visualization and isolation of antigen-stimulated primary human T cells. *Blood* **96**, 459–466 (2000).
26. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
27. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
28. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**, 550 (2014).
29. Schmidt, J. et al. NSAIDs ibuprofen, indometacin, and diclofenac do not interact with farnesoid X receptor. *Sci. Rep*. **5**, 14782 (2015).
30. Caporaso, J. G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl Acad. Sci. USA* **108** (Suppl 1), 4516–4522 (2011).
31. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
32. Edgar, R. C. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv* https://doi.org/10.1101/074161 (2016).
33. Cole, J. R. et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res*. **42**, D633–D642 (2014).
34. Whitaker, W. R., Shepherd, E. S. & Sonnenburg, J. L. Tunable expression tools enable single-cell strain distinction in the gut microbiome. *Cell* **169**, 538–546 (2017).

Acknowledgements We thank F. Gonzalez (National Institutes of Health, NIH) for providing *Nr1h4*^{fl} mice and F. Geissmann (MSKCC) for *Csf1r*^{cre} mice. We thank J. Sonnenburg (Stanford University) for providing genetic tools to engineer *B. theta*. We thank S.-E. Brown, M. Rosenthal, T. Nguyen and P. Gonzales (Boehringer Ingelheim) and R. Pinedo (Weill Cornell Medical College) for assistance with germ-free mice. We thank O. Ouerfelli and the staff at the Organic Synthesis core at MSKCC for producing 6-oxoMCA. We thank A. Pickard (MKSCC) for assistance with metabolomic data analyses. We thank E. D. D'Andrea (University of Arizona) for assistance with DSF experiments. We thank J. van der Veeken for discussion of gene-expression data and all other members of the Rudensky laboratory for suggestions and technical assistance. This study was supported by NIH grant R37, the Ludwig Institute for Cancer Research, the Hilton Foundation, and Research Beyond Borders at Boehringer Ingelheim. W.G. and C.J.G were supported by NIH grant 1DP2HD101401-01. A.Y.R is an investigator with the Howard Hughes Medical Institute.

Author contributions C.C., P.T.M and A.Y.R. conceived the study, designed experiments and wrote the manuscript; C.C. and P.T.M performed experiments and analysed data; D.K. designed and cloned the original *B. theta* constructs; K.K. and C.M. provided technical assistance with experiments; M.S. and O.I.I. analysed gene-expression data sets; J.V. maintained germ-free mouse strains; C.-J.G. and W.-B.J. analysed bacterial bile acid transformation by mass spectrometry; S.V., R.J.R. and J.R.C. quantified SCFAs by mass spectrometry; J.H. and A.Y.R. supervised the study.

Competing interests An invention disclosure has been filed based on the data generated in this study. P.T.M. and A.Y.R. received funding from Boehringer Ingelheim. A.Y.R. is a co-founder and member of the scientific advisory board of, and holds stock options in, Vedanta Biosciences. P.T.M. receives licensing royalties from Seres Therapeutics and is a co-inventor on patent applications US20170087196A1, US20180256653A1 and WO2018195467A1.

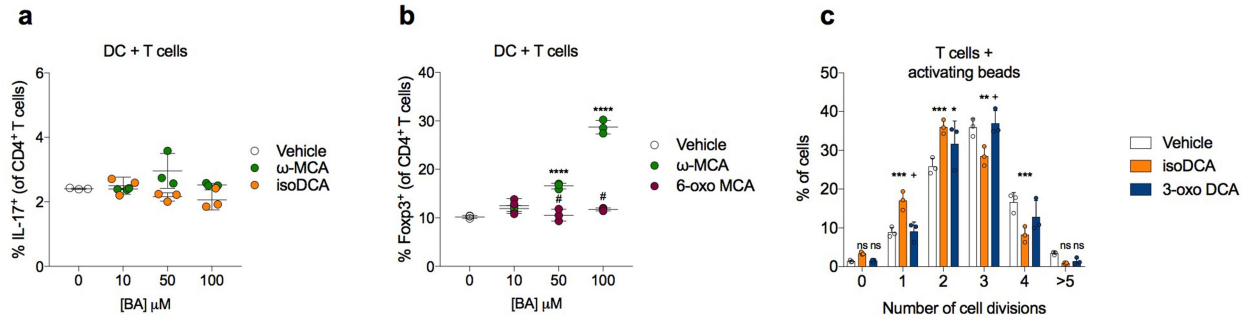
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2193-0.

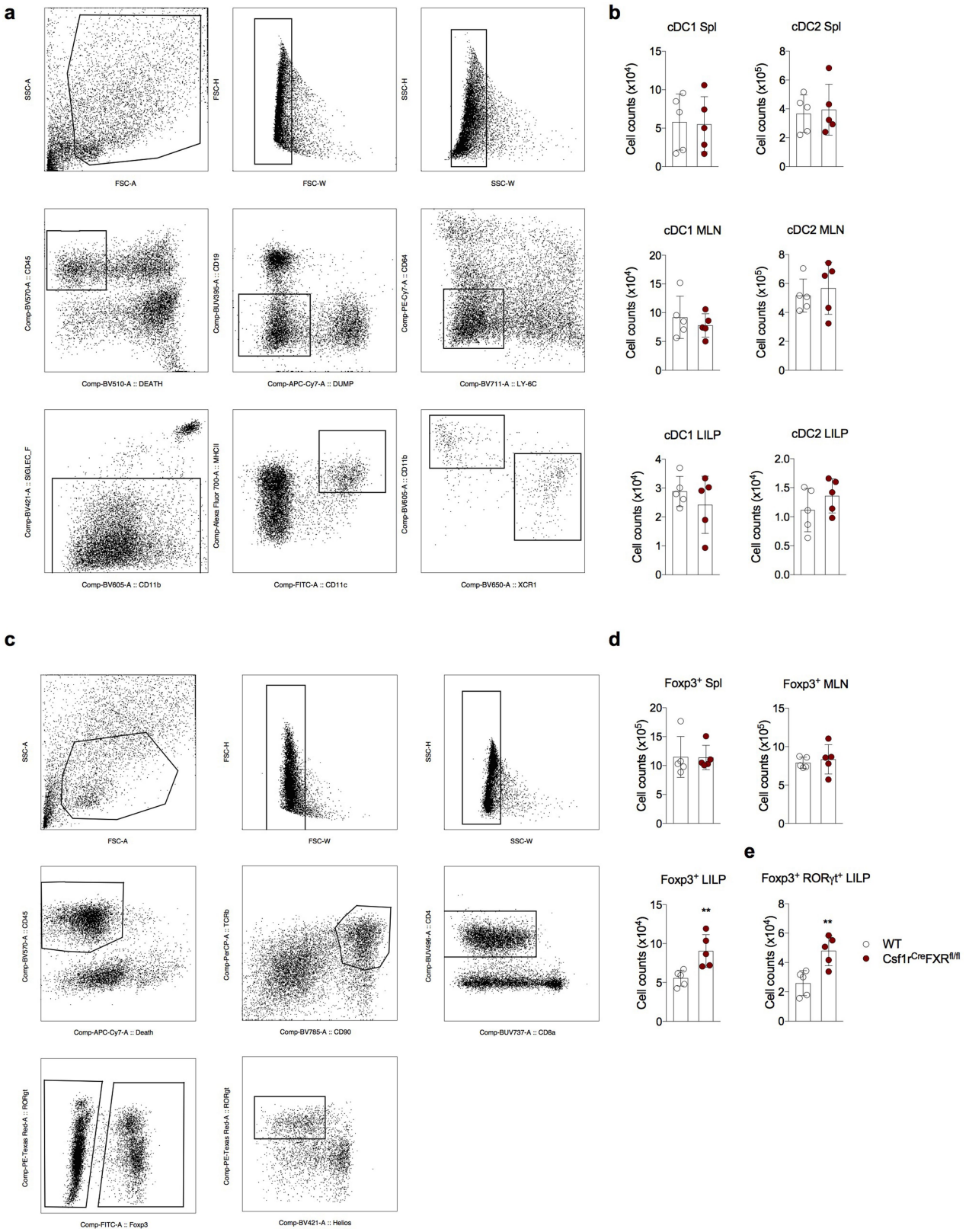
Correspondence and requests for materials should be addressed to C.C. or A.Y.R.

Peer review information *Nature* thanks Richard Steven Blumberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

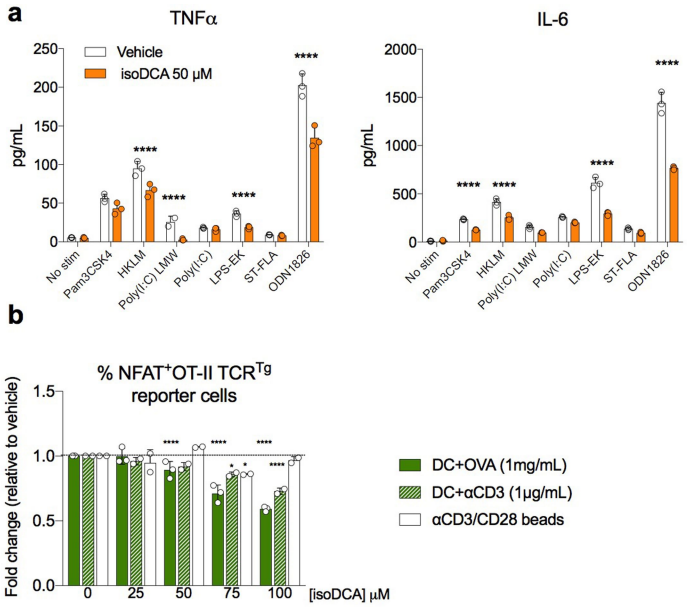
Reprints and permissions information is available at http://www.nature.com/reprints.



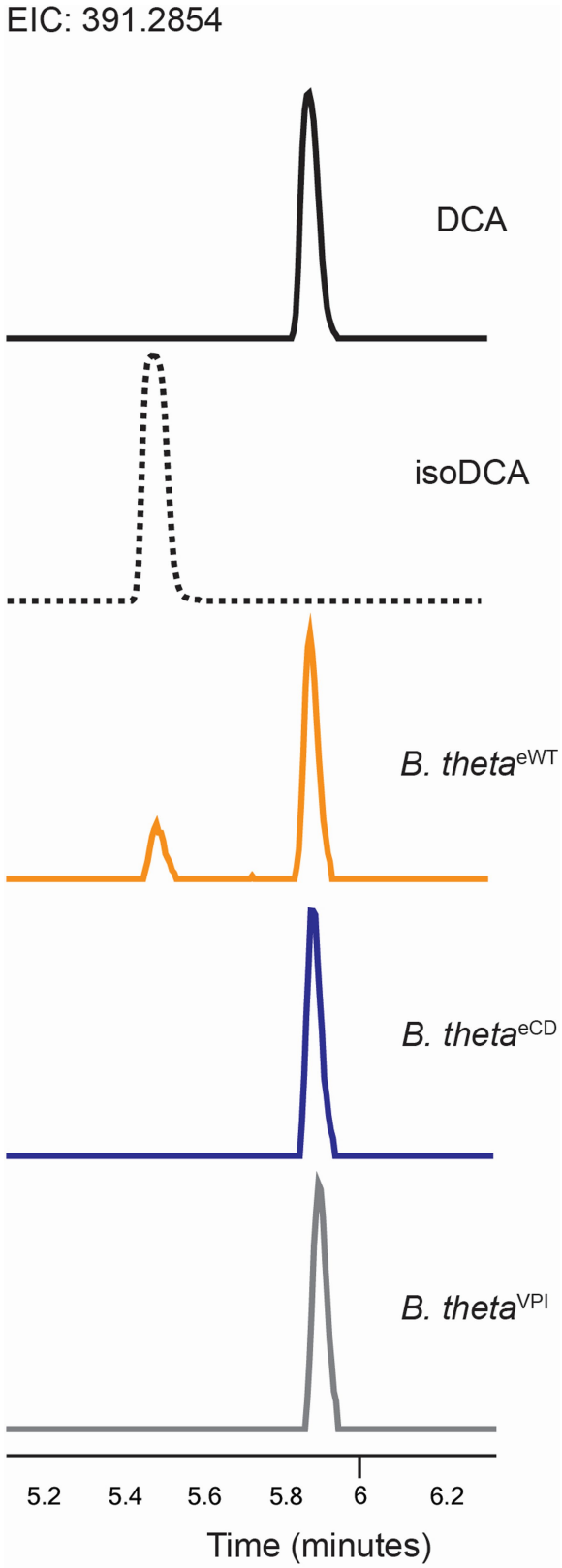
Extended Data Fig. 1 | Effects of iso- and oxo-bile acids on T cell differentiation and proliferation. a, Effects of T_{reg} cell-inducing bile acids on the in vitro generation of T_H17 cells. Naive $CD4^+$ T cells were activated by DCs in T_H17-polarizing conditions (2 ng ml^{-1} TGF-β, 1 μg ml^{-1} CD3 antibody and 20 ng ml^{-} IL-6). On day 3, cocultures were restimulated with phorbol myristate (PMA) and ionomycin in the presence of brefeldin A and monensin for 3 h before FACS analysis of IL-17 production. **b**, The 6β-OH group of ω-MCA is required for its T_{reg} cell-inducing activity. Naive $CD4^+$ T cells were activated by DCs in suboptimal T_{reg} cell-inducing conditions (1 ng ml^{-1} TGF-β, 1 μg ml^{-1} CD3 antibody and 100 U ml^{-1} IL-2) and exposed to ω-MCA or 6-oxoMCA at the indicated concentrations. Foxp3 induction was assessed by FACS on day 3. **c**, Assessment of cell division in the presence of isoDCA and 3-oxoDCA (100 μM). Naive $CD4^+$ T cells were labelled with Cell Trace Violet and activated with CD3/CD28 antibody-coated beads in the presence of TGF-β and IL-2 for three days before FACS analysis. Data shown are means ± s.d. of replicates (**a**–**c**, $n = 3$). Statistical significance determined by one-way (**a**, **b**) or two-way (**c**) ANOVA followed by a Dunnet's (**a**) or Tukey's (**b**, **c**) multiple comparison test. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ versus vehicle; hash symbol, $P < 0.05$ vs ω-MCA (paired concentration); plus symbol, $P < 0.05$ versus isoDCA (paired concentration); ns, not significant. Data are representative of at least two independent experiments.



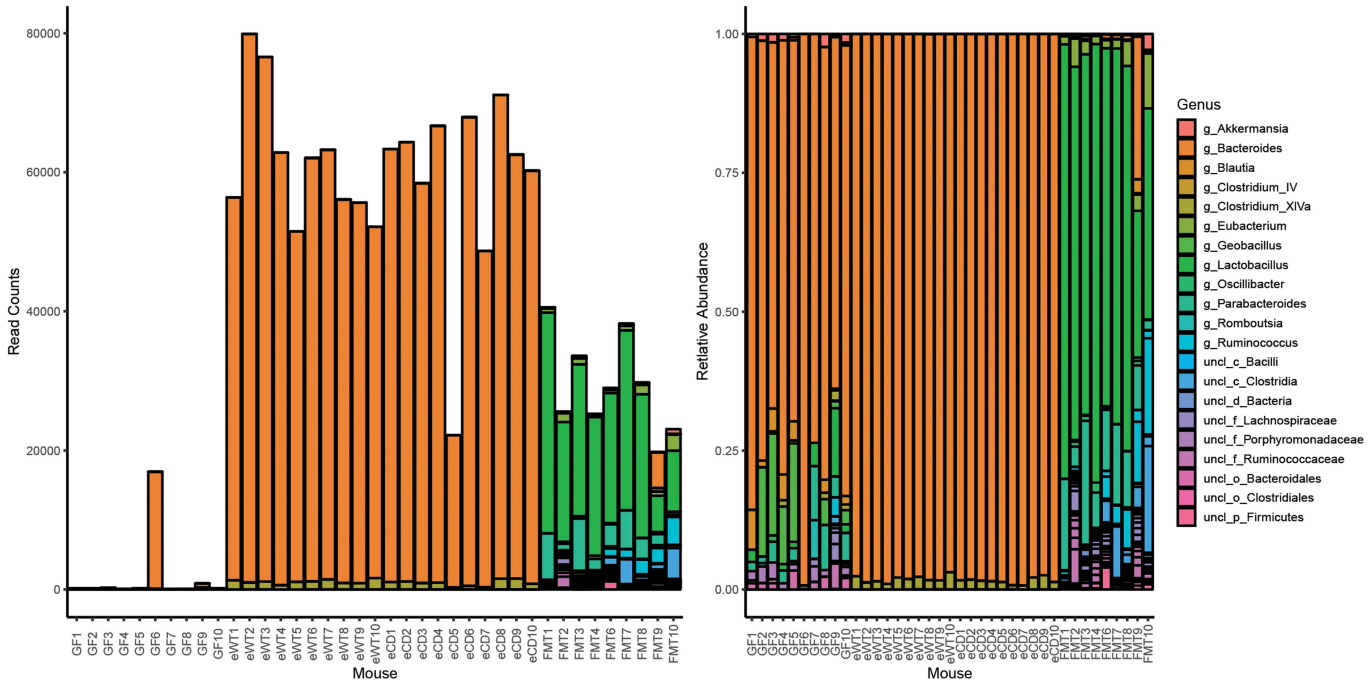
Extended Data Fig. 2 | Characterization of mice with FXR deficiency in the myeloid compartment. a–**e**, WT (*Csf1r*^{WT} *Nr1h4*$^{fl/fl}$) and DC$^{\Delta FXR}$ (*Csf1r*^{cre} *Nr1h4*$^{fl/fl}$) littermate mice were analysed between six to eight weeks of age. **a**, Gating strategy and **b**, quantification of conventional (c)DC1s (live CD45$^+$ Lin$^-$ (dump: CD90, CD3; CD64$^-$, Ly6C$^-$, Siglec-F$^-$) CD11c$^+$ MHC class II^{hi} CD11b$^-$ XCR1$^+$) and cDC2 (live CD45$^+$ Lin$^-$ (dump: CD90, CD3; CD64$^-$, Ly6C$^-$, Siglec-F$^-$) CD11c$^+$ MHC class II^{hi} CD11b$^+$ XCR1$^-$) in the spleen (Spl), mesenteric lymph node (MLN) and LILP. **c**, Gating strategy for **d**, **e**. **d**, Number of total Foxp3$^+$ T_{reg} cells in the indicated organs. **e**, Quantification of RORγt$^+$ Foxp3$^+$ T_{reg} cells in the LILP. Data shown are means ± s.d. ($n = 5$), representative of two independent cohorts of mice. Statistical significance determined by a two-tailed *t*-test. $^{**}P < 0.01$.



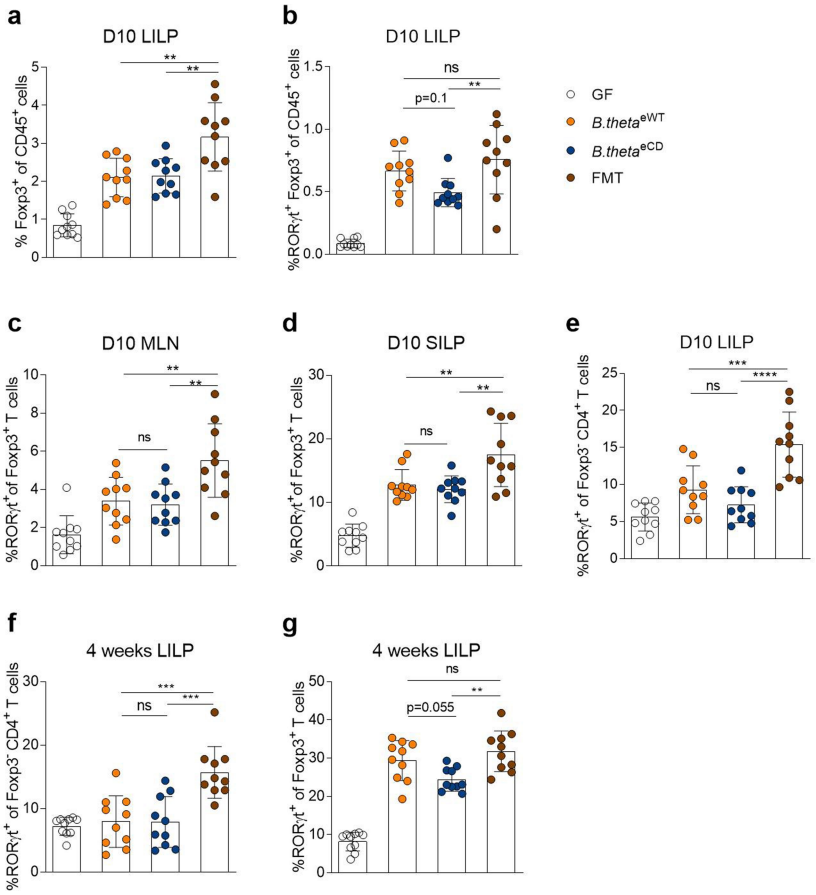
Extended Data Fig. 3 | Anti-inflammatory effects of isoDCA treatment on DCs. a, DCs (1×10^5) were stimulated for 18 h with various TLR agonists (*x*-axes) in the presence or absence of 50 μM isoDCA. Levels of the indicated cytokines (TNF-α and IL-6) in the culture supernatant were determined by ELISA. **b**, DCs (1×10^4) were pulsed with ovalbumin (OVA, 1 mg ml^{-1}) in the presence of various concentrations of isoDCA for 1 h in serum-free medium and allowed to process antigen for 4 h in complete medium before addition of an NFAT–GFP reporter cell line expressing the MHC-II-restricted OT-II TCR recognizing the ISQAVHAAHAEINEAGR peptide of OVA. The frequency of GFP$^+$ cells was determined by FACS analysis after 24 h. Cocultures treated with CD3 antibody (1 μg ml^{-1}) served as controls for DC-dependent, antigen-processing-independent effects of isoDCA on the activation of reporter cells. Activation with CD3/CD28 antibody beads in the presence of isoDCA served as a control for DC-independent effects on reporter gene expression. Shown are means ± s.d. of replicates in **a** and fold-change relative to vehicle (0 μM isoDCA) within each condition (OVA, CD3 or CD3/CD28 antibody-coated beads) in **b**. Statistical significance in **a** was determined by multiple *t*-tests using the Holm–Sidak correction method with $\alpha = 0.05$. ****$P < 0.001$ versus vehicle. Statistical significance in **b** was determined by a two-way ANOVA followed by Dunnet's multiple comparison's test. *$P < 0.05$; ****$P < 0.001$ versus vehicle in each condition. Data are representative of three independent experiments.



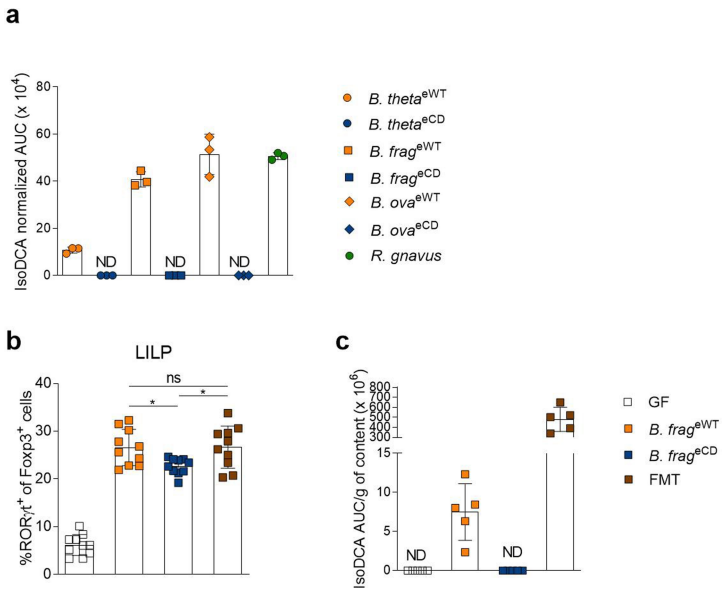
Extended Data Fig. 4 | Liquid chromatography–mass spectrometry (LC–MS)-based analysis of isoDCA production by engineered *B. theta* strains. Bacteria were grown to exponential phase and transferred to media containing DCA. Following incubation for 24 h, media was extracted with methanol and supernatants were analysed by liquid chromatography-mass spectrometry (LC–MS). Shown are traces for spike-in controls with DCA and isoDCA standards, and for media conditioned by *B. theta*^{eWT}, *B. theta*^{eCD} or the parental, unmanipulated *B. theta*^{VPI} strain VPI-5482. Data are representative of two independent experiments carried out in triplicate.



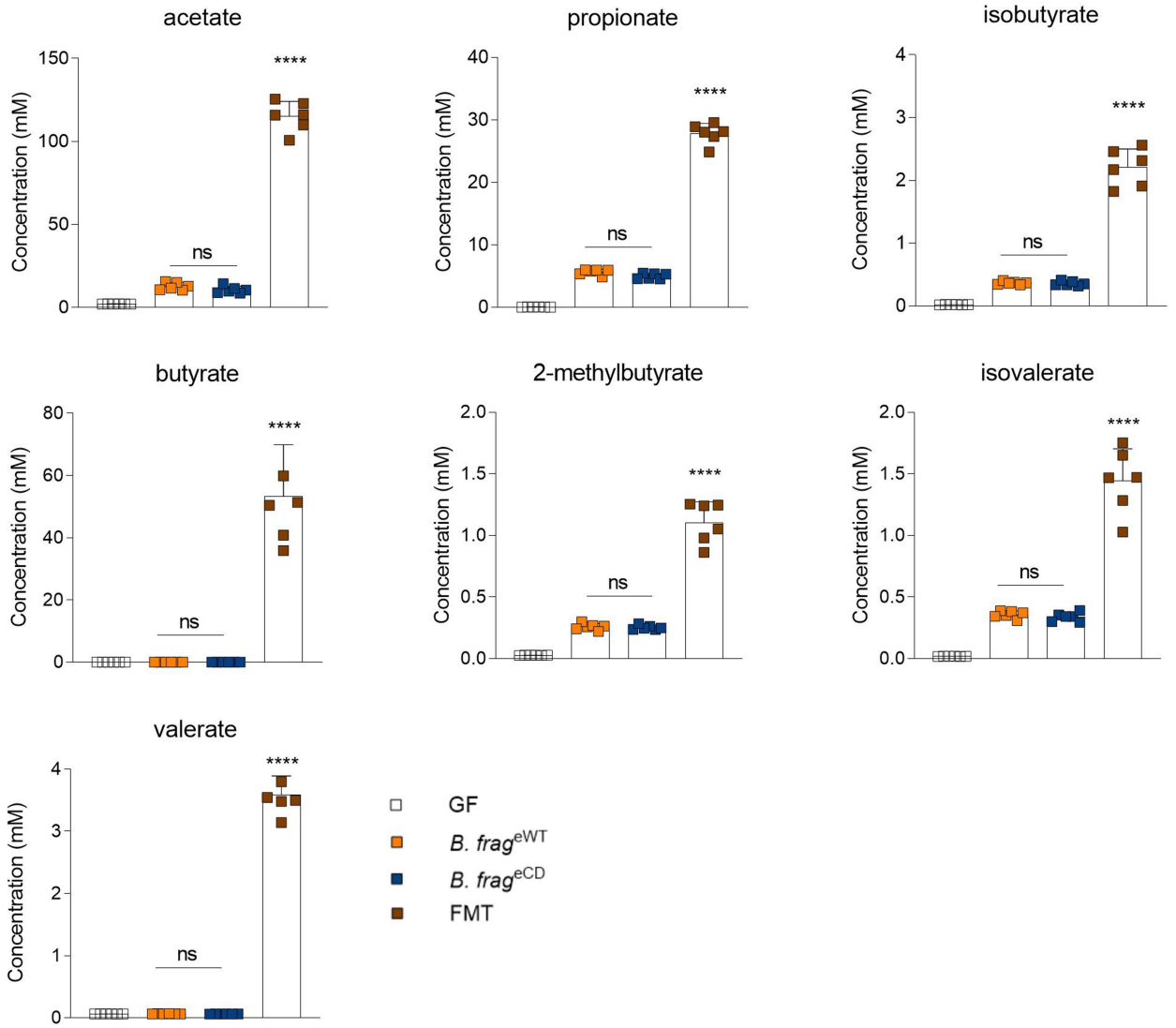
Extended Data Fig. 5 | Analyses of microbial community composition in gnotobiotic and conventionalized mice. GF mice were gavaged with WT or CD engineered consortia (*C. scindens* plus *B. theta*^{eWT} or *C. scindens* plus *B. theta*^{eCD}). Recipients of an FMT or noncolonized mice (PBS) served as references. The OTU composition of the caecal microbiota on day 10 post-colonization was determined by 16S sequencing. Shown are total read counts (left) and relative abundances (right) of bacteria in individual experimental mice, with data pooled from two independent experiments ($n = 10$).



Extended Data Fig. 6 | Effects of isoDCA-producing consortia on colonic lymphocytes. GF mice were gavaged with engineered consortia (*C. scindens* plus *B. theta*^{eWT} or *C. scindens* plus *B. theta*^{eCD}), PBS or a complex microbial community (FMT) as in Fig. 4b. **a**–**g**, Immune cell composition in the LILP was analysed by FACS on day 10 (D10; **a**–**e**) or day 30 (**f**, **g**) post-colonization. **a**, **b**, Frequencies of total Foxp3$^+$ (**a**) and RORγt$^+$ Foxp3$^+$ (**b**) T_{reg} cells among CD45$^+$ cells. **c**, **d**, Frequency of RORγt$^+$ Foxp3$^+$ cells in the MLN (**c**) and small intestine lamina propria (SILP, **d**). **e**, Frequency of RORγt$^+$ cells among Foxp3$^-$ CD4$^+$ T cells (**e**, **f**) and Foxp3$^+$ CD4$^+$ T cells (**g**). Data shown are means ± s.d. ($n = 10$), pooled from two independent experiments. Statistical significance determined by one-way ANOVA followed by Tukey's multiple comparison's test. $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$; ns, not significant.



Extended Data Fig. 7 | IsoDCA production by engineered *Bacteroides sp.* strains. a, Quantification of isoDCA production by engineered and reference strains in vitro. Bacteria were grown to exponential phase and transferred to medium containing DCA. Following incubation for 24 h, medium was extracted with methanol and supernatants were analysed by LC–MS. AUC, area under curve. **b**, **c**, GF mice were colonized with consortia containing either the engineered strain of *B. frag* capable of producing isoDCA or the catalytically dead mutant in combination with *C. scindens* (*C. scindens* plus *B. frag*^{eWT} and *C. scindens* plus *B. frag*^{eCD}, respectively). Recipients of an FMT and noncolonized mice (PBS) served as references. Immune-cell composition and isoDCA quantification were performed 10 days post-colonization. **b**, FACS analysis of the frequency of RORγt$^+$ Foxp3$^+$ CD4$^+$ T cells in the LILP. **c**, Quantification of isoDCA in caecal contents. Faecal material was weighed, homogenized and extracted with methanol for LC–MS analysis. In **a**, **c**, the AUC is normalized by the weight of the input material. Shown are means ± s.d. (**a**, $n = 3$; **b**, $n = 10$; **c**, $n = 5$). Data in **a**, **c** are representative of two independent experiments. Data in **b** are pooled from two independent experiments. Statistical significance determined by a one-way ANOVA followed by Tukey's multiple comparison's test. *$P < 0.05$; ND, not detected; ns, not significant.



Extended Data Fig. 8 | SCFA production by minimal, defined microbial consortia. GF mice were colonized with consortia containing either the engineered strain of *B. frag* capable of producing isoDCA or the catalytically dead mutant in combination with *C. scindens* (*C. scindens* plus *B. frag*^{eWT} and *C. scindens* plus *B. frag*^{eCD}, respectively). Recipients of an FMT and noncolonized mice (PBS) served as references. Caecal content material was weighed, homogenized and subjected to organic solvent extraction for GC–MS-based quantification of SCFA levels. Shown are means ± s.d. ($n = 6$), with data pooled from two independent experiments. Statistical significance determined by a one-way ANOVA followed by Tukey's multiple comparison's test. ****$P < 0.0001$; ns, not significant.

nature research

Corresponding author(s): Clarissa Campbell, Alexander Rudensky

Last updated by author(s): Feb 5, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	FACS Diva v 8.0 (BD), MiSeq Control Software v 2.6.2.1 (Illumina), MiSeq Real Time Analysis v 1.18.54 (Illumina), bcl2fastq v 2.17.1.14 (Illumina), Protein Thermal Shift v 1.3 (Applied Biosystems),
Data analysis	RNA-seq data was processed and analyzed using an in-house pipeline, available upon request: STAR v 2.4.2, 16S sequencing data was analyzed by the Microbiome Core Facility at Weill Cornell Medical College, pipeline available upon request: USEARCH v 11.0.667, Ribosome Database Project 16S training set v 16, Phyloseq v 1.30.0, FlowJo v 10.6.1, Graphpad Prism v 7.0d, MassHunter v10.1 (Agilent)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

RNA-seq data was submitted to NCBI under BioProject Accession #: PRJNA600898
16S sequencing data was submitted to NCBI under Bioproject accession #: PRJNA600979

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Sample sizes were determined by magnitude and consistency of measurable differences. Animal numbers used are given in the figure legends.
Data exclusions	No data were excluded.
Replication	Experiments are representative of 2-3 replicates with similar results, or are data pooled from 2-3 replicates as described in the figure legends. Attempts at replication were successful.
Randomization	Mice were randomly assigned to experimental groups at the beginning of treatments.
Blinding	Investigators were not blinded during group allocation and data analysis. Obvious morphological differences in the cecum of germ-free versus colonized mice make blinding difficult.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Rat monoclonal anti-mouse CD19 (BUV395) BD Biosciences Cat#563557; RRID: AB_2722495; Clone: 1D3; Dilution 1:500
Rat monoclonal anti-mouse CD45R (BUV496) BD Biosciences Cat#564662; RRID: AB_2722578; Clone: RA3-6B2; Dilution 1:500
Rat monoclonal anti-mouse CD8α (BUV737) BD Biosciences Cat#564297; RRID: AB_2722580; Clone: 53-6.7; Dilution 1:800
Rat monoclonal anti-mouse Siglec-F (BV421) BD Biosciences Cat#562681; RRID: AB_2722581; Clone: E50-2440; Dilution 1:500
Rat monoclonal anti-mouse CD45 (BV570) BioLegend Cat#103136; RRID: AB_2562612; Clone: 30-F11; Dilution 1:600
Rat monoclonal anti-mouse/human CD11b (BV605) BioLegend Cat#101237; RRID: AB_11126744; Clone: M1/70; Dilution 1:600
Mouse monoclonal anti-mouse, rat XCR1 (BV650) BioLegend Cat#148220; RRID: AB_2566410; Clone: ZET; Dilution 1:200
Rat monoclonal anti-mouse Ly-6C (BV711) BioLegend Cat#128037; RRID: AB_2562630; Clone: HK1.4; Dilution 1:800
Armenian Hamster monoclonal anti-mouse CD11c (FITC) BD Biosciences Cat#553801;RRID:AB_395060; Clone HL3; Dilution 1:400
Armenian Hamster monoclonal anti-mouse CD103 (PerCP/Cyanine5.5) BioLegend Cat#121416; RRID: AB_2128621; Clone 2E7; Dilution 1:300
Armenian Hamster monoclonal anti-mouse FcεRIα (PE) ThermoFisher Cat#12-5898; RRID: AB_466027; Clone: MAR-1; Dilution 1:500
Rat monoclonal anti-mouse F4|80 (PE-eFluor610) ThermoFisher Cat#61-4801; RRID: AB_2574612; Clone: BM8; Dilution 1:400
Rat monoclonal anti-mouse Ly-6G/Ly-6C (PE-Cy7) ThermoFisher Cat#25-5931-81; RRID: AB_469662; Clone: RB6-8C5; Dilution 1:400
Rat monoclonal anti-mouse CD117 (APC) BD Biosciences Cat#553356; RRID: AB_398536; Clone: 2B8; Dilution 1:200
Rat monoclonal anti-mouse MHC-II (redfluor710) Tonbo Biosciences Cat#80-5321; RRID: AB_2621997; Clone: M5/114.15.2; Dilution 1:800
Rat monoclonal anti-mouse CD90.2 (BV785) BioLegend Cat#105331; RRID: AB_2562900; Clone: 30-H12; Dilution 1:400

Rat monoclonal anti-mouse EpCAM (PE-Cy7) BioLegend Cat#118216; RRID: AB_1236471; Clone: G8.8; Dilution 1:400
Rat monoclonal anti-mouse CD25 (BUV395) BD Biosciences Cat#564022; RRID: AB_2722574; Clone: PC61; Dilution 1:400
Rat monoclonal anti-mouse CD4 (BUV496) BD Biosciences Cat#564667; RRID: AB_2722549; Clone: GK1.5; Dilution 1:500
Rat monoclonal anti-mouse CD62L (BV605) BioLegend Cat#104438; RRID: AB_2563058; Clone: MEL-14; Dilution 1:600
Armenian Hamster monoclonal anti-mouse Helios (Pacific Blue) BioLegend Cat#137210; RRID: AB_10575625; Clone: 22F6; Dilution 1:200
Rat monoclonal anti-mouse CD45 (BV570) BioLegend Cat#103135; RRID: AB_10898325; Clone: 30-F11; Dilution 1:1000
Rat monoclonal anti-mouse CD44 (BV650) BioLegend Cat#:103049; RRID:AB_2562600; Clone: IM7; Dilution 1:400
Rat monoclonal anti-mouse CD8a (BV711) BioLegend Cat#100759; RRID: AB_2563510; Clone: 53-6.7; Dilution 1:500
Rat monoclonal anti-mouse CD90.2 (BV785) BioLegend Cat#105331; RRID: AB_2562900; Clone: 30-H12; Dilution 1:800
Rat monoclonal anti-mouse FoxP3 (FITC) eBiosciences Cat#:11-5773-82; RRID:AB_465243; Clone: FJK-16s; Dilution 1:400
Armenian Hamster monoclonal anti-mouse TCR-beta (PerCP/Cy5.5) BioLegend Cat#:109227; RRID: AB_1575176; Clone: H57-597; Dilution 1:400
Rat monoclonal anti-mouse GITR (PE) eBiosciences Cat#12-5874-82; RRID:AB_465986; Clone:DTA-1; Dilution 1:500
Mouse monoclonal anti-mouse, RORgt (PE-CF549) BD Biosciences Cat#562684; RRID:AB_2651150; Clone:Q31-378; Dilution 1:400
Rat monoclonal anti-mouse Ep-CAM (PE/Cy7) BioLegend Cat#118216; RRID:AB_1236471; Clone:G8.8; Dilution 1:1000
Rat monoclonal anti-mouse KI-67 (AF700) BioLegend Cat#:652420; RRID: AB_2564285; Clone: 16A8; Dilution 1:600

Validation

Antibodies were not validated.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HT-29 ATCC, B16-Flt3 provided by Glenn Dranoff (Dana Farber Cancer Institute)
Authentication	Cell lines were not authenticated.
Mycoplasma contamination	Cell lines were not tested for mycoplasma.
Commonly misidentified lines (See ICLAC register)	Cell lines were not identified in ICLAC.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Germ Free C57Bl/6 mice were purchased from Taconic. Female mice were used in all experiments to facilitate distribution of animals into experimental groups. Germ Free Foxp3GFPΔCNS1 and Foxp3GFP mice were bred and maintained at Weill Cornell Medical College Gnotobiotics facility. Male and female littermate mice were used in all experiments. Animals were fed with autoclaved 5KA1 chow. Germ Free status was routinely checked by aerobic and anaerobic cultures of fecal samples for bacteria and fungi and by PCR of fecal DNA samples for bacterial 16S and fungal/yeast 18S genes. GF and ex-GF mice were at least 8 weeks old at the initiation of experiments. SPF mice were housed at the Research Animal Resource Center for Memorial Sloan Kettering Cancer Center and Weill Cornell Medical College with 12-hour light/dark cycles under ambient conditions and ad libitum access to food and water. Experimental mice were maintained in a standard rodent diet (5053, LabDiet). Dr. Frank Gonzalez (NIH, USA) kindly provided the Nr1h4floxed mouse strain. Csf1rCre mice were provided by Dr. Frederic Geissman at MSKCC, USA. CD4Cre mice [Tg(Cd4-cre)1Cwi] were purchased from Jax laboratories and maintained in house. Experimental littermate animals were generated by mating mice homozygous for the Nr1h4floxed allele, with one of the breeders (male of female) carrying one copy of the Cre-driver gene. Cells from male and female CD4CreNr1h4floxed or Csf1rCreNr1h4floxed were used for in vitro experiments. Male Csf1rCreNr1h4floxed mice were analyzed at 6-8 weeks of age.
Wild animals	This study did not use wild animals.
Field-collected samples	This study did not use field-collected samples.
Ethics oversight	Studies were approved by the IACUCs of Memorial Sloan Kettering Cancer Center and Boehringer Ingelheim Pharmaceuticals, Inc.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	This information is included in the methods.
Instrument	Flow Cytometry: LSR II (BD Biosciences), Cell Sorting: Aria II (BD Biosciences)
Software	FACS Diva v8.0 (BD Biosciences), FlowJo v10.6.1
Cell population abundance	Purification by sorting was performed using FACSAria II cell sorter (BD Biosciences), with > 98% Purity
Gating strategy	Gating strategy is described in the methods section.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

Origin of complexity in haemoglobin evolution

https://doi.org/10.1038/s41586-020-2292-y

Received: 30 July 2019

Accepted: 7 April 2020

Published online: 20 May 2020

Check for updates

Arvind S. Pillai[1], Shane A. Chandler[2,8], Yang Liu[3,8], Anthony V. Signore[4,8], Carlos R. Cortez-Romero[5], Justin L. P. Benesch[2], Arthur Laganowsky[3], Jay F. Storz[4], Georg K. A. Hochberg[1,7] & Joseph W. Thornton[1,6]✉

Most proteins associate into multimeric complexes with specific architectures[1,2], which often have functional properties such as cooperative ligand binding or allosteric regulation[3]. No detailed knowledge is available about how any multimer and its functions arose during evolution. Here we use ancestral protein reconstruction and biophysical assays to elucidate the origins of vertebrate haemoglobin, a heterotetramer of paralogous α- and β-subunits that mediates respiratory oxygen transport and exchange by cooperatively binding oxygen with moderate affinity. We show that modern haemoglobin evolved from an ancient monomer and characterize the historical 'missing link' through which the modern tetramer evolved—a noncooperative homodimer with high oxygen affinity that existed before the gene duplication that generated distinct α- and β-subunits. Reintroducing just two post-duplication historical substitutions into the ancestral protein is sufficient to cause strong tetramerization by creating favourable contacts with more ancient residues on the opposing subunit. These surface substitutions markedly reduce oxygen affinity and even confer cooperativity, because an ancient linkage between the oxygen binding site and the multimerization interface was already an intrinsic feature of the protein's structure. Our findings establish that evolution can produce new complex molecular structures and functions via simple genetic mechanisms that recruit existing biophysical features into higher-level architectures.

The interfaces that hold molecular complexes together typically involve sterically tight, electrostatically complementary interactions among many amino acids[4]. Similarly, allostery and cooperativity usually depend on numerous residues that connect surfaces to active sites[5]. The acquisition of such complicated machinery would seem to require elaborate evolutionary pathways. The classical explanation of this process, by analogy to the evolution of morphological complexity, is that multimerization conferred or enhanced beneficial functions, allowing selection to drive the many substitutions required to build and optimize new interfaces[4,6].

Whether this account accurately describes the evolution of any natural molecular complex requires a detailed reconstruction of the historical steps by which it evolved. Haemoglobin (Hb) is a useful model for this purpose, because the structural mechanisms that mediate its multimeric assembly, cooperative oxygen binding, and allosteric regulation are well established[7,8]. Moreover, its subunits descend by duplication and divergence from the same ancestral proteins, so their history can be reconstructed in a single analysis. Despite considerable speculation[9–11], virtually nothing is known about the evolutionary origin of Hb's heterotetrameric architecture and the functions that depend on it.

From monomer to homodimer

We inferred the phylogeny of Hb and closely related globins (Fig. 1a, Extended Data Fig. 1a, b, e). The duplication that produced the distinct Hbα and Hbβ subunits occurred before the last common ancestor of jawed vertebrates (Fig. 1a). The closest outgroups—myoglobin (Mb)[12], globin E[13], and globin Y (Extended Data Fig. 1d)—are monomers. A more distant clade of agnathan 'haemoglobin' and vertebrate cytoglobin includes monomers and dimers[14,15], but the dimers assemble through interfaces that differ from each other and from those used in Hb, indicating parallel acquisition[16,17]. These observations suggest that the Hb $\alpha_2\beta_2$ heterotetramer evolved from an ancestral monomer via an unknown intermediate form.

To characterize when and how the tetramer evolved, we first reconstructed Hb of the ancestral jawed vertebrate by phylogenetically inferring the maximum a posteriori sequences of the ancestral α- and β-subunits (Ancα and Ancβ; Fig. 1a, Extended Data Fig. 1b, c). We coexpressed and purified Ancα and Ancβ and characterized their assembly using native mass spectrometry (nMS), size-exclusion chromatography (SEC) and multi-angle light scattering (MALS). Like extant Hb, Ancα and Ancβ associate into $\alpha_2\beta_2$ heterotetramers, with a tetramer–dimer

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [2]Department of Chemistry, Chemistry Research Laboratory, University of Oxford, Oxford, UK. [3]Department of Chemistry, Texas A&M University, College Station, TX, USA. [4]School of Biological Sciences, University of Nebraska, Lincoln, NE, USA. [5]Program in Cell and Molecular Biology, University of Chicago, Chicago, IL, USA. [6]Department of Human Genetics, University of Chicago, Chicago, IL, USA. [7]Present address: Max Planck Institute for Terrestrial Microbiology, Marburg, Germany. [8]These authors contributed equally: Shane A. Chandler, Yang Liu, Anthony V. Signore. ✉e-mail: joet1@uchicago.edu

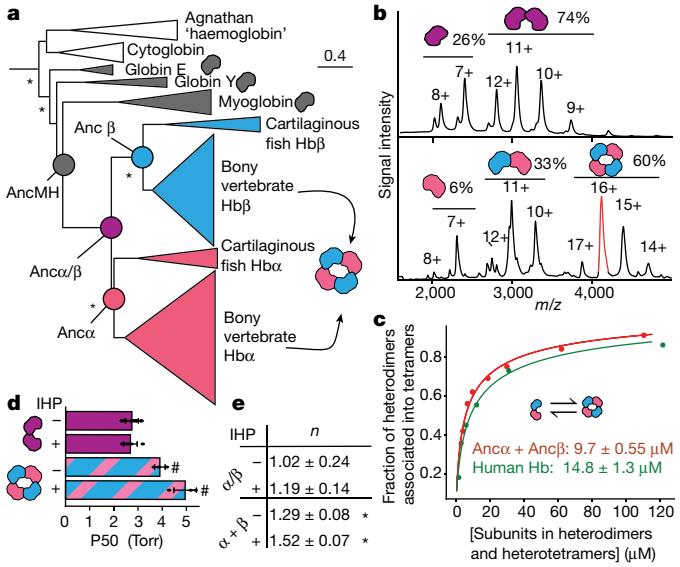


Fig. 1 | Structure and function of ancestral globins. a, Simplified phylogeny of vertebrate globins. Icons, oligomeric states. *Approximate likelihood ratio statistic >10. Complete phylogeny in Extended Data Fig. 1a. Circles, reconstructed ancestral proteins. Scale bar, substitutions per site. **b**, nMS spectra of Ancα/β (top, purple) and Ancα + Ancβ (lower, pink and blue) at 20 μM. Charge states, stoichiometries, and occupancy (fraction of moles of subunits) shown. Red, analysed by tandem mass spectrometry (MS/MS) in Extended Data Fig. 2e. **c**, Dimer-to-tetramer affinity of Ancα + Ancβ (red) and human Hb (green). Circles, fraction of α + β heterodimers incorporated into $\alpha_2\beta_2$ tetramers, measured once by nMS. K_d ± s.e. (in moles of subunits) estimated by nonlinear regression. **d**, **e**, Oxygen affinity (P50) and cooperativity (Hill coefficient, *n*) of Ancα/β and Ancα + Ancβ. IHP, twofold molar excess of inositol hexaphosphate. Mean ± 95% confidence interval (CI) from 3–5 replicates (dots) shown. #, P50 significantly different from Ancα/β under corresponding conditions ($P < 0.01$, *t*-test). *Significant cooperativity ($n \neq 1$, $P < 0.05$, *F*-test; Extended Data Fig. 1f).

dissociation constant (K_d) of 10 μM, comparable to that of human Hb (15 μM; Fig. 1b, c, Extended Data Fig. 2a–c, f, i). Expressed in isolation, Ancα forms homodimers (Extended Data Fig. 3a), and Ancβ forms homotetramers (Extended Data Fig. 3b), just as extant Hb subunits do[18,19]. The heterotetrameric structure of Hb therefore evolved before the jawed vertebrate ancestor, more than 400 million years ago.

By contrast, Ancα/β, the pre-duplication ancestral protein, homodimerizes with a K_d of 9 μM measured by nMS, but does not form tetramers (Fig. 1b, Extended Data Fig. 2d, f, g). Even at 1.4 mM, no tetramers are detectable using SEC (Extended Data Fig. 2h). Ancα/β was therefore a homodimer, with virtually no propensity to tetramerize. This result is robust even when we incorporate statistical uncertainty about the ancestral sequence in several alternative constructs (Extended Data Fig. 4). This is also the most parsimonious history, because extant Hbα dimerizes and Hbβ tetramerizes when they are expressed in isolation[18,19]: a monomeric Ancα/β would imply independent gains of dimerization, and a tetramer would require early gain of tetramerization followed by loss in Hbα (Extended Data Fig. 4).

AncMH, the common ancestor of Hb and myoglobin, is monomeric. No higher-order stoichiometries were detected using nMS of His-tagged AncMH at 70 μM (Extended Data Fig. 3f). Even at 600 μM, only monomers were apparent using SEC (Extended Data Fig. 2j). The untagged protein also does not dimerize at concentrations at which Ancα/β is predominantly dimeric, as shown using SEC and a globin-specific concentration assay on lysate from transformed cells (Extended Data Fig. 3d, e). A monomeric AncMH is also the most parsimonious scenario, because its closest outgroups are all monomers (Extended Data Fig. 4b–e).

The Ancα/β homodimer is therefore the evolutionary missing link between an ancient monomer and the Hb heterotetramer. After duplication, a novel interaction evolved, enabling these dimers to associate into tetramers.

Evolution of Hb functions

We characterized the evolution of the functional properties of Hb by assaying the oxygen-binding characteristics of the ancestral proteins. The physiological role of modern Hb—loading oxygen in the lungs or gills and unloading it in the periphery—is possible because Hb binds and releases oxygen cooperatively and has an affinity lower than that of myoglobin; its affinity is further reduced by allosteric effectors[8]. Like human Hb, the coexpressed and copurified complex Ancα + Ancβ displays measurable cooperativity, and its oxygen affinity is similar to that of stripped, recombinant human Hb[20] (Fig. 1d, e). The affinity of Ancα + Ancβ is reduced in the presence of the allosteric effector inositol hexaphosphate (IHP), although by less than that of human Hb[20]. The functional characteristics of extant Hb were therefore in place by the jawed vertebrate ancestor.

By contrast, the oxygen affinity of Ancα/β is significantly higher than that of Ancα + Ancβ, and it does not display detectable cooperativity or allosteric regulation by IHP (Fig. 1d, e, Supplementary Discussion). The major functional characteristics of modern Hb therefore evolved between Ancα/β and Ancα + Ancβ, the same interval during which tetramerization evolved. This also represents the most parsimonious history: Hb tetramers are cooperative, but Hbα homodimers and Hbβ homotetramers are not[18,21], suggesting that this property did not yet exist in their common ancestor (Extended Data Fig. 4).

Because Ancα/β lacked cooperativity, allostery, or reduced affinity, it could not have performed modern Hb's physiological role in oxygen exchange. Furthermore, the first step in the evolution of Hb's tetrameric architecture—acquisition of homodimerization from a monomeric ancestor—could not have been driven by selection for the major functional properties of Hb, because the homodimer did not possess any of them.

Ancestral and derived interfaces

Hb assembles via two distinct interfaces on each subunit: IF1 mediates α1–β1 and α2–β2 contacts, while IF2 mediates α1–β2 and α2–β1 contacts[7] (Fig. 2a). To determine which interface evolved before Ancα/β, we applied hydrogen–deuterium exchange mass spectrometry (HDX-MS) to Ancα/β. We compared patterns of deuterium uptake at high and low protein concentrations (at which dimers or monomers predominate, respectively; Extended Data Fig. 2d, f, g). Solvent-exposed residues incorporate deuterium faster than buried residues, so peptides that contribute to the dimer interface should exhibit higher deuterium uptake when the monomeric state predominates. We found that Ancα/β peptides containing residues in IF1 incorporate significantly more deuterium under monomer-favouring than dimer-favouring conditions; no difference was observed for IF2 (Fig. 2b, c, Extended Data Figs. 5–8). Moreover, mutations in residues in IF1 substantially impair dimerization of Ancα/β, but a mutation that disrupts IF2 in human Hb[22] has no effect (Fig. 2d, Extended Data Figs. 7c, 9). Reverting all IF1 residues in Ancα/β to the amino acid state from AncMH yields predominantly monomers, but reverting those at IF2 has no effect (Fig. 2d, Extended Data Fig. 7d).

Ancα/β homodimers therefore assembled via IF1. After duplication, IF2 evolved, enabling dimers to assemble into tetramers (Fig. 2e). Corroborating this inference, extant Hbα homodimers assemble via IF1, whereas Hbβ tetramers use both IF1 and IF2, indicating that IF1 was inherited from their ancestor Ancα/β[18,19]. The finding that IF2 evolved after the gene duplication explains why Ancα/β is neither cooperative nor allosterically regulated, because both functions require IF2-mediated assembly into tetramers[23].

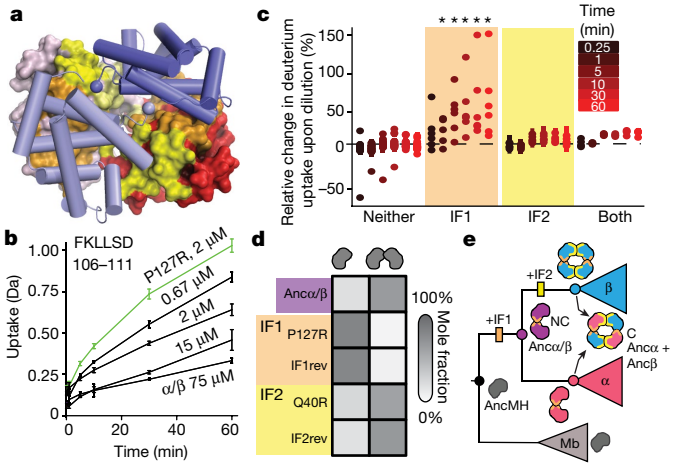


Fig. 2 | Identification of homodimerization interface in Ancα/β. a, Hb heterotetramers assemble via two interfaces (IF1, orange; IF2, yellow) on each subunit. Red and pink surfaces, α-subunits; blue cartoon, β-subunits. Ancα + Ancβ homology model is shown. **b**, Deuterium incorporation by an Ancα/β peptide that contributes to IF1 (Extended Data Fig. 5g, h). Uptake (mean ± s.e. from three replicates per incubation time) is shown for Ancα/β (black) and monomeric IF1 mutant P127R (green). **c**, Each circle shows mean difference in deuterium uptake by one Ancα/β peptide when expressed at monomer-favouring versus dimer-favouring concentrations (0.67 and 75 μM, respectively; three replicates each, with s.e.). Peptides are classified by the interface to which they contribute and coloured by incubation time. *Mean uptake in interface category significantly different from other categories ($P < 0.05$, permutation test, Extended Data Figs. 6g, 7). **d**, Dimer and monomer occupancy by Ancα/β and mutants, assessed using nMS at 20 μM. P127R and Q40R disrupt IF1 and IF2, respectively. IF1rev and IF2rev revert historical substitutions to state in AncMH (spectra in Extended Data Fig. 7c, d). **e**, Evolution of Hb tetramer. Rectangles, acquisition of IF1 and IF2. C, cooperative; NC, noncooperative; Mb, myoglobin.

Genetic mechanisms for the new interface

The causal substitutions for the evolution of heterotetramers from the homodimer must have occurred on one or both of the post-duplication branches that lead from Ancα/β to Ancα and Ancβ. On the Ancα branch, there were only three changes, of which none were at IF2. On the Ancβ branch, there were 42 changes, including 5 at IF2 and 4 others at IF1 (Fig. 3a).

We introduced the IF2 substitutions into Ancα/β (Ancα/β5) and found that they confer strong assembly into tetramers when Ancα/β5 is coexpressed with Ancα; the mixture includes both heterotetramers and homotetramers (Fig. 3b, Extended Data Fig. 10c, d). A version containing only four of the substitutions (Ancα/β4) formed homotetramers at 20 μM but did not heteromerize with Ancα. The fifth change (H104E) therefore confers the capacity to associate with Ancα, probably because it interacts with His104 on Ancα, forming a hydrogen bond in the heteromer but clashing in the homomer (Fig. 3b, Extended Data Fig. 10a, b). Even a subset of just two IF2 changes (Ancα/β2) causes high-affinity assembly into homotetramers (K_d = 1 μM; Fig. 3b, d, Extended Data Fig. 10g). The genetic basis for the evolution of a new strong interface was therefore simple.

The IF2 substitutions are not sufficient to yield specific occupancy of the $\alpha_2\beta_2$ architecture: coexpressing Ancα/β5 with Ancα produces a mixture of tetramers containing zero, one, or two α-subunits (Fig. 3b, Extended Data Fig. 10c, d). We hypothesized that IF1 substitutions conferred heterospecificity by favouring the assembly of heterodimers across IF1, which then form $\alpha_2\beta_2$ heterotetramers across IF2. We introduced the IF1 substitutions into Ancα/β5 (Ancα/β9) and coexpressed it with Ancα. As predicted, heterotetramers and heterodimers predominate over homomers (Fig. 3c). Ancα/β9 + Ancα is poorly soluble, preventing quantification by nMS, but the addition of five historical

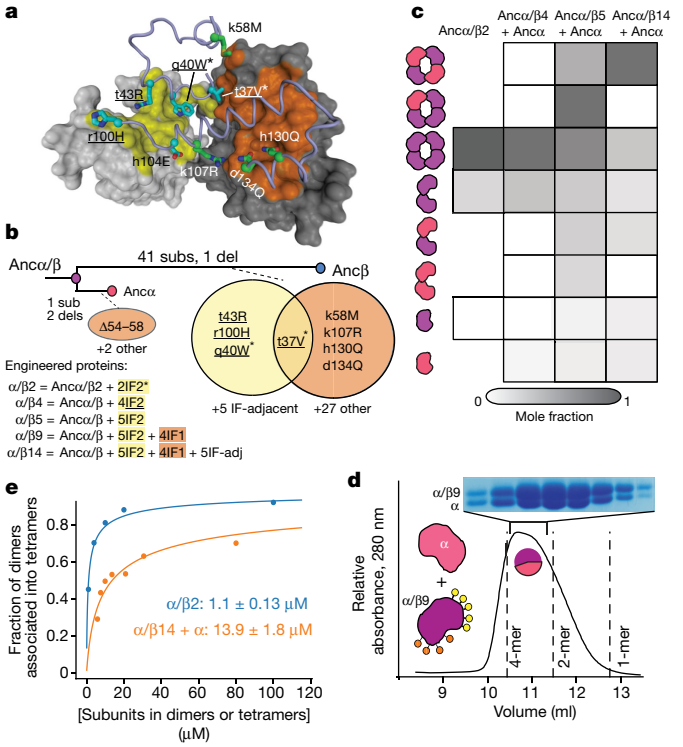


Fig. 3 | Genetic mechanisms of tetramer evolution. a, Homology model of Ancα + Ancβ tetramer with interface residues substituted between Ancα/β and Ancβ. Grey surfaces, two Ancα subunits; yellow, IF2; orange, IF1. Blue cartoon, partial backbone of one Ancβ subunit; sticks, side chains of substituted sites (cyan, IF2; green, IF1). Labels show state in Ancα/β (lowercase) and Ancβ (uppercase). *Sites in Ancα/β2; underlined, sites in Ancα/β4. **b**, Phylogenetic interval between Ancα/β and Ancα + Ancβ with number of substitutions and deletions per branch. Venn diagrams, sites substituted at interfaces. Below, substitutions incorporated in mutant proteins. **c**, Occupancy of multimers measured by nMS at 20 μM, as fraction of moles of subunits in each state. Ancα/β2 was expressed in isolation, so only homomers are plotted. Spectra in Extended Data Fig. 10. **d**, SEC of Ancα/β9 + Ancα at 80 μM. Lines, elution volumes of tetramer (Ancα + Ancβ), dimer (Ancα/β), monomer (human Mb). Pie chart, proportions of Ancα and Ancα/β9 subunits in tetramer-containing fraction, measured by denaturing MS (Extended Data Fig. 11e). Top, electrophoresis of tetramer-containing fraction. **e**, Dimer-to-tetramer affinity of Ancα/β2 (blue) and Ancα/β14 + Ancα (orange). Orange circles, fraction of Ancα/β14 + Ancα heterodimers incorporated into heterotetramers; blue, fraction of Ancα/β2 homodimers in homotetramers, measured by nMS once. K_d ± s.e. estimated by nonlinear regression.

substitutions at sites proximal to the interfaces (Ancα/β14 + Ancα) improves solubility, and nMS confirmed preferential occupancy of $\alpha_2\beta_2$ heterotetramers (K_d = 6 μM; Fig. 3b, d, Extended Data Fig. 10e, f).

The Hb heterotetramer therefore evolved from the Ancα/β homodimer via two sets of substitutions. Changes at IF2 created a strong new interface that conferred tetramerization; changes at IF1 yielded heterospecificity. In both cases, only a few substitutions were required.

Structural mechanisms for the new interface

We next investigated how so few substitutions could have generated a new and specific multimeric interaction. Using a homology model of the heterotetramer, we identified all favourable contacts that mediate association across the ancestral interfaces and used the phylogeny to determine when these amino acids evolved (Fig. 4a–c, Extended Data Fig. 10h, i).

The substitutions that conferred tetramerization recruited residues that already existed on the opposing surface into newly favourable interactions. All 13 residues that Ancα contributes to IF2 are unchanged from

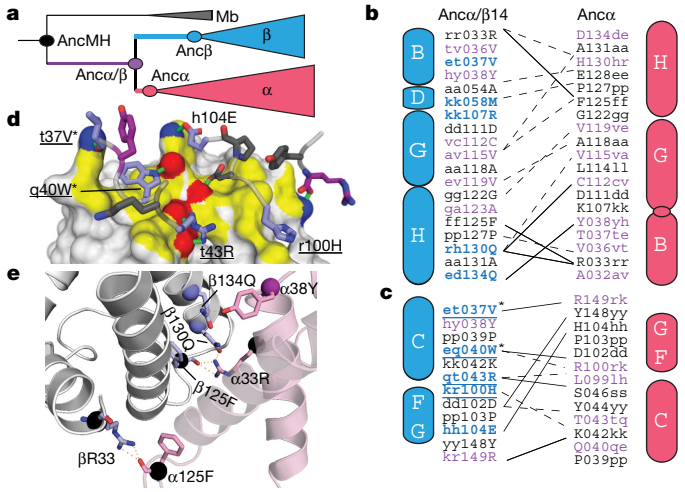


Fig. 4 | Structural mechanisms of evolution of Hb interfaces. a, Phylogenetic classification of ancestral states and substitutions. Black, state in AncMH; purple, substituted from AncMH to Ancα/β; blue or red, substituted from Ancα/β to Ancβ or Ancα. **b**, **c**, Contact maps for residues buried at IF1 (**b**) and IF2 (**c**) of Ancα + Ancβ. Residues coloured by scheme in **a**. Letters show state in AncMH (outside, lower case), Ancα/β (middle, lower case) and Ancβ or Ancα (inside, upper case). Solid lines, predicted hydrogen bonds; dashed lines, van der Waals interactions. Underlined, substitutions in Ancα/β4; *substitutions in Ancα/β2. Cylinders, helices (Extended Data Fig. 2a). Oval, deletion of helix D in Ancα. **d**, IF2 contacts in Ancα + Ancβ. Grey surface, Ancα, with yellow IF2; hydrogen-bonding atoms are red (oxygen) or blue (nitrogen), with bonds as green lines. Cartoon, Ancβ backbone, with IF2-interacting sidechains (sticks, coloured as in **a**). **e**, Close-up of IF1 in Ancα + Ancβ model. Sticks, hydrogen-bonding residues; spheres, Cα atoms; coloured as in **a**.

their ancestral state in Ancα/β, and many were acquired earlier (Fig. 4c). The IF2 substitutions on the Ancβ branch yielded new van der Waals contacts and hydrogen bonds with these ancient residues (Fig. 4c, d). For example, the ring of Trp40 (substituted in Ancβ from the ancestral glutamine) nestles tightly in an ancient hydrophobic indentation on Ancα. Similarly, the IF1 substitutions that increase occupancy of the $\alpha_2\beta_2$ heterotetramer all modify interactions with ancient residues that were conserved on Ancα (Fig. 4b, e).

Both interfaces also involve favourable contacts between residues that were unchanged from their deep ancestral states in both subunits. In IF1, for example, R33 on each subunit donates two hydrogen bonds to F125 on the facing surface, and both residues evolved before AncMH. Each subunit contains both residues, and IF1 occurs twice in the tetramer, so these two sites form a total of eight hydrogen bonds in the complex (Fig. 4b, e). Similarly, IF2 contains several hydrogen bonds and Van der Waals interactions between pairs of residues that originated before Ancα/β.

Because of the exponential relationship between binding energy and affinity, one substitution can markedly increase the occupancy of the multimer, if it builds on the foundation of even very weak interactions between older residues. Satisfying an unpaired hydrogen-bond donor or acceptor or burying a hydrophobic ring can contribute up to 16 kJ mol^{-1} to an association[24,25]. Each interface occurs twice in Hb (Fig. 2a), so a substitution that confers a favourable interaction does so twice in the tetramer, doubling its effect on binding free energy and reducing K_d by up to six orders of magnitude. A single mutation can therefore shift occupancy of the tetramer from virtually nonexistent to the predominant species.

Mechanisms of cooperativity

Finally, we sought insight into the evolution of the cooperativity and reduced affinity of Ancα + Ancβ. Cooperativity in extant Hb involves two conformational states that all subunits can adopt: one has higher affinity for oxygen but weaker IF2 contacts between subunits than the other[23,26]. Cooperativity is classically thought to be mediated by an 'allosteric core'—the set of residues on the helix that connect the haem to IF2, which is positioned differently in the two conformations[27].

To understand the mechanisms that triggered the evolution of cooperativity and reduced oxygen affinity, we first examined the phylogenetic history of residues in the haem pocket and allosteric core. At sites within 4 Å of the haem, no substitutions occurred during the interval when cooperativity was acquired. The vast majority were acquired before AncMH (Fig. 5a, Extended Data Fig. 1c), including the proximal histidine, which covalently binds the haem iron and transduces the movement of the haem upon oxygen binding to the allosteric core and IF2, thereby causing the other subunits to shift between low- and high-affinity conformations. Two substitutions occurred in Ancβ on the helix that connects IF2 to the histidine, but there were none in Ancα (Fig. 5a), and both subunits make the conformational transition in extant Hb. These observations suggest that the structural properties that mediate the allosteric linkage between the haem–oxygen-binding site and IF2 already existed in Ancα/β, before cooperativity and tetramerization evolved. Consistent with this idea, many of the conformational changes that mediate Hb cooperativity, such as distortion of the haem's geometry and movement of the histidine and helix upon oxygen binding, also occur in myoglobin, which is monomeric and noncooperative[28,29].

We hypothesized that, because of this ancient structural connection between the IF2 surface and the active site, evolution of the intersubunit interaction across IF2 was sufficient to confer cooperativity and reduce affinity. We characterized oxygen binding by Ancα/β2, which contains only two historical substitutions at IF2. As predicted, we found that these mutations reduce oxygen affinity by two- to threefold compared to Ancα/β (Fig. 5b); they also confer weak but statistically significant cooperativity (Extended Data Fig. 5b). Acquisition of the tetrameric association alone therefore changes the oxygen-binding function of the protein and confers cooperative oxygen binding.

The tetramer's ability to transition between high- and low-affinity states, however, is sensitive to mutation. Ancα/β4 and the Ancα/β14 + Ancα heterotetramer also have reduced oxygen affinity relative to Ancα/β, but they lose the cooperativity found in Ancα/β2 (Fig. 5b). A likely explanation is that the additional mutations in these constructs overstabilize the low-affinity conformation relative to the high-affinity state. If so, then some of the other substitutions that occurred between Ancα/β and the cooperative complex Ancα + Ancβ must have tuned this equilibrium so that both conformations can be occupied, depending on the oxygen partial pressure (Fig. 5c). The order in which these changes occurred cannot be resolved: the IF2 substitutions may have immediately generated a cooperative Hb-like complex, similar to Ancα/β2; alternatively, cooperativity may have evolved via a low-affinity tetrameric intermediate, like Ancα/β4 (Fig. 5c).

Evolution of molecular complexity

Our findings establish that a few genetic changes drove the evolution of Hb's complex structure and functions from its dimeric precursor. Other molecular complexes may also have evolved by short mutational paths. Interactions between proteins and other kinds of substrates, such as DNA or small molecules, have historically evolved via one or a few historical substitutions[30], and we see no reason why multimeric interactions in general should be more difficult to evolve. Multimers can be engineered from non-assembling precursors by one or a few mutations[31,32], and naturally occurring point mutations are known to cause disease by inducing higher-order complexes[33].

The simple mechanism by which Hb appears to have evolved its cooperativity—acquisition of binding to a molecular partner at a new interface—could explain the origin of cooperativity and allostery in

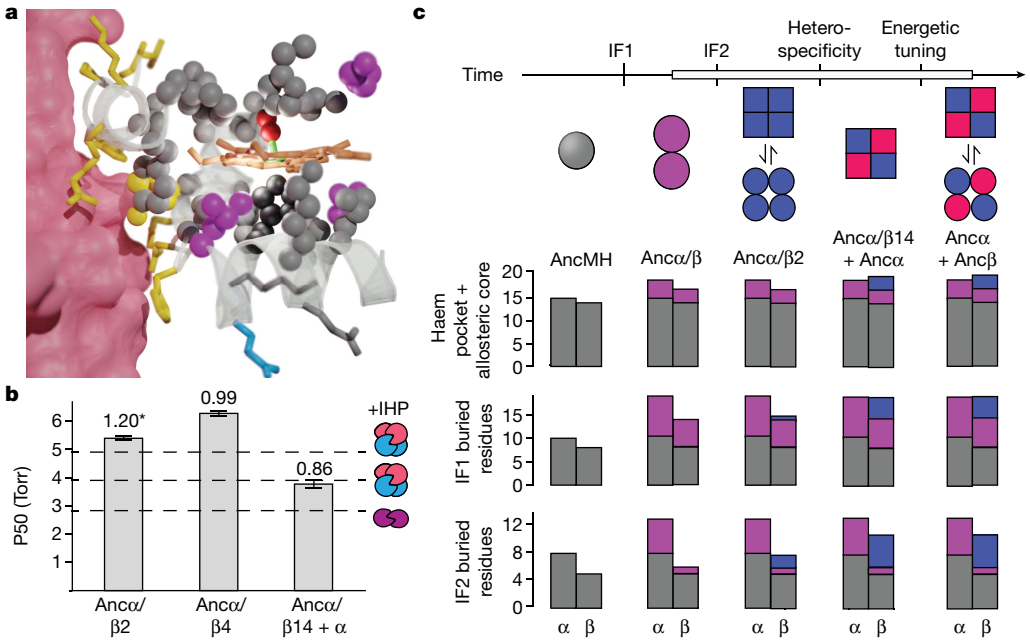


Fig. 5 | Evolution of cooperativity by interface acquisition. a, Haem pocket and IF2 in Ancα + Ancβ. Pink surface, one Ancα. Tan sticks, haem (with green iron and red oxygen). Spheres, Ancβ residues within 4 Å of haem, coloured by temporal category: grey, conserved since AncMH (dark grey, iron-coordinating histidine); purple, conserved since Ancα/β; blue, substituted between Ancα/β and Ancβ. Sticks, other residues on helix connecting histidine to IF2, coloured temporally. Yellow, Ancβ residues at IF2. No changes near haem or IF2 occurred in Ancα. **b**, Oxygen binding by Ancα/β mutants with historical substitutions. P50 ± SE, with Hill coefficient *n* above, estimated by nonlinear regression under effector-stripped conditions (raw data in Extended Data Fig. 10j). *Significant cooperativity ($n \neq 1$, $P \leq 0.05$, *F*-test, Extended Data Fig. 1f). Dotted lines, affinities of Ancα + Ancβ and Ancα/β, which is unaffected by IHP. **c**, Top, evolution of the cooperative Hb heterotetramer. Circles and squares, conformations with high and low oxygen affinity, respectively. Two IF2 substitutions cause homotetramerization, cooperativity, and reduced affinity (**b**). Other substitutions that confer heterotetramerization change the relative stabilities of high and low-affinity conformations, abolishing or restoring cooperativity. White box, interval in which order of substitutions is unknown. Bottom, acquisition of residues in structurally defined categories in Ancα and Ancβ, coloured by temporal category. No changed occurred in Ancα.

other systems[34,35]. If two plausible conditions are met—the new interface is near or structurally connected to the functionally active site, and the optimal conformation for binding is different from the optimal conformation for activity—then binding will impair activity, and vice versa. Given this tradeoff, the evolution of binding alone will confer cooperativity or negative allostery.

The history of Hb shows that complex molecular structures and functions can arise by means other than the long, gradual trajectories of functional optimization by which biological complexity has long been thought to evolve[6,36]. In principle, molecular assemblies could arise and become more complex via neutral processes[37–39], but this scenario is unlikely if many mutations are required. Our work shows that the higher-level multimeric state and functional properties of Hb evolved through just a few mutations, which fortuitously built upon and interacted with ancient structural features. These older features could not have been initially acquired because of selection for the functions of the final complex, because they existed before those functions first appeared. Some are likely to have originated and been preserved by selection for more ancient functions, while others may have appeared transiently by chance. Although evolution of any particular molecular sequence or architecture without consistent selection for those properties is vanishingly improbable, our findings suggest that proteins evolve constantly through a dense space of possibilities in which complex new interactions and functional states are easily accessible.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2292-y.

1. Ahnert, S. E., Marsh, J. A., Hernández, H., Robinson, C. V. & Teichmann, S. A. Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).
2. Marsh, J. A. & Teichmann, S. A. Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.* **84**, 551–575 (2015).
3. Monod, J., Wyman, J. & Changeux, J. P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88–118 (1965).
4. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).
5. Rivalta, I. et al. Allosteric pathways in imidazole glycerol phosphate synthase. *Proc. Natl Acad. Sci. USA* **109**, E1428–E1436 (2012).
6. Dawkins, R. *Climbing Mount Improbable* (WW Norton & Company, 1997).
7. Perutz, M. F. et al. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature* **185**, 416–422 (1960).
8. Storz, J. F. *Hemoglobin: Insights into Protein Structure, Function, and Evolution* (Oxford Univ. Press, 2018).
9. Goodman, M. & Moore, G. W. Phylogeny of hemoglobin. *Syst. Zool.* **22**, 508–532 (1973).
10. Coates, M. L. Hemoglobin function in the vertebrates: an evolutionary model. *J. Mol. Evol.* **6**, 285–307 (1975).
11. Zuckerkandl, E. The evolution of hemoglobin. *Sci. Am.* **212**, 110–118 (1965).
12. Kendrew, J. C. et al. Structure of myoglobin: a three-dimensional Fourier synthesis at 2 A. resolution. *Nature* **185**, 422–427 (1960).
13. Blank, M. et al. Oxygen supply from the bird's eye perspective: globin E is a respiratory protein in the chicken retina. *J. Biol. Chem.* **286**, 26507–26515 (2011).
14. Fago, A., Rohlfing, K., Petersen, E. E., Jendroszek, A. & Burmester, T. Functional diversification of sea lamprey globins in evolution and development. *Biochim. Biophys. Acta. Proteins Proteomics* **1866**, 283–291 (2018).
15. Lechauve, C. et al. Cytoglobin conformations and disulfide bond formation. *FEBS J.* **277**, 2696–2704 (2010).
16. Heaslet, H. A. & Royer, W. E., Jr. The 2.7 A crystal structure of deoxygenated hemoglobin from the sea lamprey (*Petromyzon marinus*): structural basis for a lowered oxygen affinity and Bohr effect. *Structure* **7**, 517–526 (1999).
17. Makino, M. et al. High-resolution structure of human cytoglobin: identification of extra N- and C-termini and a new dimerization mode. *Acta Crystallogr. D* **62**, 671–677 (2006).

18. Kidd, R. D., Baker, H. M., Mathews, A. J., Brittain, T. & Baker, E. N. Oligomerization and ligand binding in a homotetrameric hemoglobin: two high-resolution crystal structures of hemoglobin Bart's (γ(4)), a marker for α-thalassemia. *Protein Sci.* **10**, 1739–1749 (2001).
19. Kumar, K. K., Jacques, D. A., Guss, J. M. & Gell, D. A. The structure of α-haemoglobin in complex with a haemoglobin-binding domain from *Staphylococcus aureus* reveals the elusive α-haemoglobin dimerization interface. *Acta Crystallogr. F* **70**, 1032–1037 (2014).
20. Hoffman, S. J. et al. Expression of fully functional tetrameric human hemoglobin in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **87**, 8521–8525 (1990).
21. Tyuma, I., Benesch, R. E. & Benesch, R. The preparation and properties of the isolated α and β subunits of hemoglobin A. *Biochemistry* **5**, 2957–2962 (1966).
22. Manning, L. R., Dumoulin, A., Jenkins, W. T., Winslow, R. M. & Manning, J. M. Determining subunit dissociation constants in natural and recombinant proteins. *Methods Enzymol.* **306**, 113–129 (1999).
23. Ackers, G. K. Energetics of subunit assembly and ligand binding in human hemoglobin. *Biophys. J.* **32**, 331–346 (1980).
24. Fersht, A. R. et al. Hydrogen bonding and biological specificity analysed by protein engineering. *Nature* **314**, 235–238 (1985).
25. Eisenberg, D. & McLachlan, A. D. Solvation energy in protein folding and binding. *Nature* **319**, 199–203 (1986).
26. Mihailescu, M.-R. & Russu, I. M. A signature of the T → R transition in human hemoglobin. *Proc. Natl Acad. Sci. USA* **98**, 3773–3777 (2001).
27. Gelin, B. R., Lee, A. W. M. & Karplus, M. Hemoglobin tertiary structural change on ligand binding. Its role in the co-operative mechanism. *J. Mol. Biol.* **171**, 489–559 (1983).
28. Sato, A., Gao, Y., Kitagawa, T. & Mizutani, Y. Primary protein response after ligand photodissociation in carbonmonoxy myoglobin. *Proc. Natl Acad. Sci. USA* **104**, 9627–9632 (2007).
29. Barends, T. R. M. et al. Direct observation of ultrafast collective motions in CO myoglobin upon ligand dissociation. *Science* **350**, 445–450 (2015).
30. Siddiq, M. A., Hochberg, G. K. & Thornton, J. W. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* **47**, 113–122 (2017).
31. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548**, 244–247 (2017).
32. Grueninger, D. et al. Designed protein-protein association. *Science* **319**, 206–210 (2008).
33. Pauling, L. et al. Sickle cell anemia, a molecular disease. *Science* **110**, 543–548 (1949).
34. Coyle, S. M., Flores, J. & Lim, W. A. Exploitation of latent allostery enables the evolution of new modes of MAP kinase regulation. *Cell* **154**, 875–887 (2013).
35. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **147**, 1564–1575 (2011).
36. Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* 204–208 (John Murray, 1859).
37. Lynch, M. Evolutionary diversification of the multimeric states of proteins. *Proc. Natl Acad. Sci. USA* **110**, E2821–E2828 (2013).
38. Finnigan, G. C., Hanson-Smith, V., Stevens, T. H. & Thornton, J. W. Evolution of increased complexity in a molecular machine. *Nature* **481**, 360–364 (2012).
39. Gray, M. W., Lukeš, J., Archibald, J. M., Keeling, P. J. & Doolittle, W. F. Irremediable complexity? *Science* **330**, 920–921 (2010).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Sequence data and alignment

We collected 177 annotated amino acid sequences of haemoglobin and related paralogues in 72 species from UniPROT, Ensembl and NCBI RefSeq. Sequences were aligned using MAFFT v7[40]. The maximum likelihood (ML) phylogeny and branch lengths were inferred from the alignment using PHYML v3.1[41] and the LG model[42] with gamma-distributed among-site rate variation and empirical state frequencies. This best-fit evolutionary model was selected using the Akaike Information Criterion in PROTTEST. Node support was evaluated using the approximate likelihood ratio test statistic (aLRS), which expresses the difference in likelihood between the most likely topology and the most likely topology that does not include the split of interest; aLRS has been shown to be reasonably accurate, robust, and efficient compared to other means of characterizing support[43,44]. The tree was rooted on neuroglobin and globin X, paralogues that are found in both deuterostomes and protostomes[45]. Tetrapods possess three Hbα paralogues, Hbα^{A}, Hbα^{D}, and Hbα^{Z} (also known as Hbα^{E} or Hb-zeta)[46]. The ML phylogeny inferred from this alignment contained a weakly supported sister relationship between all Actinopterygian Hbα genes and the tetrapod Hbα^{Z}, to the exclusion of tetrapod Hbα^{A} and Hbα^{D}. This is a nonparsimonious scenario, because it requires an early gene duplication and subsequent loss of the Hbα^{A}/Hbα^{D} lineage in Actinopterygii. We therefore constrained the topology to unite tetrapod HbαA^{A}, Hbα^{D}, and Hbα^{Z} in a clade (Extended Data Fig. 1a). PhyML v3.1 was then used to re-infer the best-fit branch topology and branch lengths given this constraint.

Ancestral sequences were reconstructed and the posterior probability distributions of ancestral states were inferred using the ML method using the codeml package in PAML 4.9[47], given the ML-constrained phylogeny and branch lengths. Historical substitutions were assigned to phylogenetic branches as differences between the maximum a posteriori amino acid states between parent and daughter nodes. The asymmetry between the branch lengths leading from Ancα/β to Ancα and to Ancβ has been observed previously[48] and presumably reflects there being more shared amino acid states between Hbα and the outgroups (myoglobin, globins E and Y, and so on) than between Hbβ and the outgroups. The sequences for reconstructed ancestors have been deposited in GenBank (IDs MT079112, MT079113, MT079114, MT079115).

Recombinant protein expression

Ancestral genes were codon-optimized for expression in *Escherichi coli* using CodonOpt and generated by de novo DNA synthesis (IDT gBlocks). For globin expression, coding sequences were cloned into a pLIC expression vector without affinity tags and expressed under a T7 polymerase promoter. For oxygen-affinity measurements, plasmid pCOMAP[49], which expresses *E. coli* methionine aminopeptidase 1 (MAP1), was cotransformed to ensure efficient N-terminal methionine excision. For co-expression of two globins, sequences were expressed from a polycistronic operon in plasmid pGM, without tags and under a T7 promoter, separated by a spacer containing a stop codon and ribosome binding site. *E. coli* methionine aminopeptidase 1 (MAP1) was coexpressed from the same plasmid.

JM109 DE3 *E. coli* cells (NEB) were transformed and plated into solid Luria broth (LB) medium containing 50 μg/ml carbenicillin (and 50 μg/ml kanamycin, if pCOMAP was being cotransformed). A single colony was inoculated into 50 ml LB with appropriate antibiotics and grown overnight. Five millilitres of this culture was inoculated into a larger 500-ml LB culture. Cells were grown at 37 °C and shaken at 225 rpm in an incubator (New Brunswick 126) until they reached an optical density at 600 nm (OD600) of 0.4–0.6. The culture was then supplemented with 0.5 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG) and 50 mg/l hemin (Sigma). After 4 h of expression at 37 °C, CO was bubbled through the solution for 10 min and cells were collected by centrifugation at 5,000*g*. Protein purification was carried out immediately after expression.

Protein purification by ion exchange

Ancα/β, P127R, V119A, Ancα/β4 + Ancα and the alternative ancestral reconstructions were purified using ion exchange chromatography[20,49]. All buffers were saturated with CO before purification and vacuum filtered through a 0.2 μM PFTE membrane (Omnipore) to remove particulates. After expression, cells were resuspended in 200 ml of 50 mM Tris (pH 6.8) with 2 cOMPLETE protease inhibitor tablets (Roche) and 0.5 mM DTT. The cell suspension was lysed in 50-ml batches in a glass beaker using an FB505 sonicator with a power setting of 90%, 1 s on/off for 2 min. The lysate was then centrifuged at 30,000*g* to eliminate cell debris, inclusion bodies and aggregates. The supernatant was further syringe-filtered used HPX Millex Durapore filters (Millipore). A HiTrap SP cation exchange (GE) column was attached to an FPLC system (AKTAprime plus) and equilibrated in 50 mM Tris (pH 6.8). Lysate was passed over the column. The SP column was washed with 200 ml of 50 mM Tris to eliminate weakly bound contaminants. Bound Hbs eluted with a 100-ml gradient of 50 mM Tris (pH 6.9) 1 M NaCl, from 0 mM to 1 M. Fractions (0.5 ml) were collected along the length of the gradient. The four reddest fractions were collected and then concentrated in an Amicon μLtra-15 tube by centrifugation at 4,000*g* to a final volume of 500 μl. The sample was injected into a Sephacryl Hiprep 16/60 S-100 HR size-exclusion column (SEC) for additional purification. The column was equilibrated in phosphate buffered saline (PBS) at pH 7.4. Depending on molecular weight, purified globins elute at 48–52 ml (tetramer), 56–60 ml (dimer) or 64–67 ml (monomer). The purity and identity of isolated proteins was assessed using 20% SDS–PAGE and denaturing HRA-MS. The purified proteins were concentrated and then flash frozen with liquid nitrogen until use.

Protein purification by zinc affinity chromatography

Ancα/β5 + Ancα, Ancα/β9 + Ancα, Ancα/β14 + Ancα, and Ancα + Ancβ were purified using zinc-affinity chromatography, adapted from a published method[50]. Buffers were loaded onto the metal affinity column using an AKTAprime FPLC. To prepare the zinc affinity column, nickel was removed from a HisTrap column (GE) using stripping buffer (100 mM EDTA, 100 mM NaCl, 20 mM TRIS, pH 8.0). The column was then washed with diH_2O for five column volumes. Then 0.1 M $ZnSo_4$ was passed over the column until conductance reached a stable value. The column was then washed with five column volumes of water. After expression, cells were resuspended in 50 ml lysis buffer containing 20 mM Tris and 150 mM Nacl (pH 7.4). The cells were sonicated as described above. The lysate was passed over a zinc-affinity HisTrap column. The column was washed with 200 ml wash buffer (20 mM Tris and 150 mM Nacl, pH 7.4). The bound Hbs were eluted with a 50-ml gradient of imidazole, up to 500 mM, and 0.5-ml fractions were collected during the run. The four reddest fractions were collected. The Hb-containing fractions were concentrated and injected into a Sephacryl S-100 HR column for additional purification, as described above.

Purification of globin Y

The globin Y sequences of *Callorhincus milli* (NCBI reference sequence NP_001279719.1) and *Xenopus laevis* (NCBI reference sequence NP_001089155.1) were synthesized (IDT, Coralville, IA, USA) and cloned into a pLIC vector with an N-terminal hexahistidine tag (MHHHHHH). Expression and lysis were carried out under the same conditions as described above. The bacterial lysate was passed over a 5-ml HisTrap nickel-affinity column (GE). The column was washed with five column volumes of wash buffer (20 mM Tris and 150 mM Nacl, pH 7.4). The bound globins were eluted with a 15-ml gradient of imidazole from 0 to 500 mM; five fractions of equal volume were collected. The three reddest fractions were combined. The eluted protein was concentrated to 2 ml, passed through a 0.45-μm filter, and subjected to a final purification

by SEC using a Sephacryl S-100 HR column and an AKTA Prime FPLC system. Globin Y eluted in fractions collected between 61 and 64 ml.

Purification of his-tagged AncMH

The sequence of AncMH was codon-optimized for expression in *E. coli*, synthesized, and cloned into a pLIC vector with an N-terminal hexahistidine tag, because untagged AncMH was not readily purifiable. Recombinant expression, cell lysis, and purification were carried out under the conditions described for globin Y.

Characterization of protein stability

Protein stability was measured by circular dichroism (CD) using a JASCO 1500 CD spectrophotometer. Experiments were conducted at protein concentration of 10 μM (50 mM sodium fluoride, 20 mM sodium phosphate buffer) in a 0.2-mm path length quartz cell. CD spectra were collected at 2 °C intervals (10 min each) as the temperature was increased from 25 °C to 95 °C. Molar ellipticity at 222 nm was measured four times at each temperature; the mean was then divided by the value of molar ellipticity at 222 nm at room temperature (25 °C) to estimate the fraction of unfolded protein. To estimate the melting point (T_m) of each protein, a custom script was written to find the best fit parameters (T_m and slope) for the Boltzmann sigmoid function: fraction unfolded = $1/(1+e^{(t-Tm)/slope)})$. All three ancestral proteins were stable, with $T_m > 60$ °C (Extended Data Fig. 1c).

High-resolution denaturing mass spectrometry

Two hundred microlitres of purified protein was placed in a Slide-A-Lyzer MINI dialysis unit that was suspended in 500 ml of 50 mM ammonium acetate. The solution was stirred overnight at 4 °C. After dialysis, the proteins were transferred to a microfuge tube and centrifuged at 30,000*g* to eliminate aggregates. The concentration was adjusted to 20 μM. Half a microlitre of sample was sprayed using an Agilent 6224 Tof Mass Spectrometer at fragment voltage 200 V. Protein masses were estimated by maximum entropy mass deconvolution implemented in MassHunter (Agilent).

Size-exclusion chromatography and multi-angle light scattering

All proteins were converted to the CO-bound form by adding sodium dithionite to 5 mg/ml, desalting on a Sephadex G-25 desalting column equilibrated with CO-saturated PBS (150 mM NaCl, pH 7.4), and then passing CO through the eluent. Protein concentration was measured by UV absorbance at 280 nm (Tryptophan) and 419 nm (HbCO-specific) using a Nanodrop 2000c (Thermo-scientific). For analytic SEC, a Superdex 75 10/300 GL column (GE) was equilibrated in CO-saturated PBS, and then injected with 500 μl sample, using a 500-μl injection loop on an AKTAprime and monitored by absorbance at 280 nm. For SEC coupled with MALS, a Superdex 200 10/300 GL column was injected with 150 μl sample on the AKTAprime; light scattering and refractive index of eluent were measured using a Dawn Helios-II (Wyatt) light scattering detector and Optilab T-rEX refractometer, respectively. Molar mass fitting was carried out using Astra software.

Globin concentration assay

After protein expression, cells harvested by centrifugation from one 500-ml culture were resuspended in 15 ml PBS and sonicated as described above. Cell debris and aggregate were removed by centrifugation at 20,000*g*. Remaining lysate was concentrated to 5 ml in Amicon μLtra-15 centrifuge concentrators (3,000 NMWL). Five hundred microlitres of this sample was injected into a superdex-75 10/300 GL column. Fractions of eluent (0.2 ml) were collected. We took 50 μl from each fraction and added it to 150 μl Hemoglobin Assay kit reagent (Sigma) in one well of a 96-well plate. In each plate, 50 μl of a 100 mg/dl calibrator (Sigma) was also added to 150 μl of Hemoglobin Assay kit reagent (Sigma) in one well. We used 50 μl PBS added to the 150 μL reagent as a blank. Absorbance was measured at 400 nm using a Victor x5 plate reader (PerkinElmer). Haem concentration in each fraction was measured using the following equation: concentration = $62.5 \times (OD_{sample} - OD_{blank})/(OD_{calibrator} - OD_{blank})$ μM.

Oxygen affinity and cooperativity

Purified proteins were deoxygenated using sodium dithionite at 10 mg/ml and immediately passed through a PD-10 desalting column (GE Healthcare) equilibrated with 25 ml of 0.01 M HEPES/0.5 mM EDTA (pH 7.4). Eluted proteins were concentrated using Amicon μLtra-4 Centrifugal Filter Units (Millipore). Equilibrium oxygen-binding assays were performed at 25 °C using a Blood Oxygen Binding System (Loligo Systems), using 0.1 mM protein (haem concentration) dialysed in 0.1 M HEPES/0.5 mM EDTA buffer. Protein solution was sequentially equilibrated at 3–5 different oxygen tensions (PO_2) yielding 30–70% saturation while continually monitoring absorbance at 430 nm (deoxy peak) and 421 nm (oxy/deoxy isosbestic point). Plots of fractional saturation against PO_2 were constructed from these measurements, and the Hill equation was fit to each plot using OriginPro 2016, yielding estimates of P50 (PO_2 at half-saturation) and the cooperativity coefficient (n, the slope at half saturation in the Hill plot, n50). We collected 95% Cis on parameter estimates by multiplying the s.e.m. over replicate experiments by 1.96 (Fig. 1d, e). The statistical significance of cooperativity was assessed by using an F-test to compare the fit of the data to a model in which n is a free parameter to a null model in which $n = 1$.

To assess the potential for ancestral proteins to have been regulated by allosteric effectors, assays were performed in stripped medium or with IHP added at 0.5 mM. Although IHP may not have been the physiological effector in ancestral organisms, it has been shown to allosterically regulate Hbs from representatives of all major vertebrate lineages, whereas other organic phosphates such as 2,3-biphosphoglycerate (BPG), ATP, and GTP have more lineage-specific effects[51–53]. IHP therefore serves as a useful 'all-purpose' polyanion to test the allosteric regulatory capacity of the ancestral Hb. There is ample precedent for using IHP to study Hb allostery irrespective of whether it is the authentic physiological effector[50,54–56]. This is because IHP modulates $Hb–O_2$ affinity in a manner that is qualitatively similar to those of other effectors, including BPG, ATP, GTP, and IPP[8,50]. These molecules all share the same mechanism of action, reversibly binding a set of cationic residues in the cleft between β_1 and β_2 subunits, and thereby stabilizing the low-affinity T conformation via electrostatic interactions[57–59].

Native mass spectrometry

Proteins were buffer exchanged into 200 mM ammonium acetate with a centrifugal desalting column (Micro Bio-Spin P-6, BioRad) and loaded into a gold-coated glass capillary. Samples were ionized for MS measurement by electrospray ionization. MS and MS/MS ion isolation were performed on a Synapt G1 HDMS instrument (Waters Corporation) equipped with a radio frequency generator to isolate higher m/z species (up to 32k) in the quadrupole, and a temperature-controlled source chamber as previously described[60]. Instrument parameters were tuned to maximize signal intensity for MS and MS/MS while preserving the solution state of the protein complexes. All samples were sprayed at room temperature. Instrument settings were: source temperature of 50 °C, capillary voltage of 1.7 kV, sampling cone voltage of 100 V, extractor cone voltage of 5 V, trap collision energy of 25 V, argon flow rate in the trap set to 7 ml/min (5.6×10^{-2} mbar), and transfer collision energy set to 15 V. The T-wave settings were for trap (300 ms^{-1}/1.0 V), IMS (300 ms^{-1}/20 V) and transfer (100 ms^{-1}/10 V), and trap DC bias (30 V). For MS/MS, ion isolation was achieved using the same settings as described above, with the quadrupole LM resolution set to 6. Activation of protein complexes for individual monomer identification was achieved by increasing the trap collision voltage to 120 V in MS/MS mode, with all other settings unchanged. Analysis of the MS and MS/MS data to estimate masses and relative abundances was performed with the software program Unidec[61].

The occupancy of each stoichiometric state was calculated as the proportion of globin subunits in that state, based on the summed areas under the corresponding peaks in the spectrum. To estimate K_d of the monomer-to-dimer transition Ancα/β, we performed nMS at variable protein concentrations. At each concentration, the observed fraction of subunits incorporated into dimers (F_d) was estimated as $F_d = 2x_d/(x_m + 2x_d)$, where x_m and x_d are the sums of the signal intensities of all peaks corresponding to the monomeric and dimeric stoichiometries, respectively. This procedure was repeated at a range of protein concentrations. Nonlinear regression was then used to find the best-fit value of K_d using the equation

$$F_d = \frac{1}{P_{tot}} \times \frac{(4P_{tot} + K_d) - \sqrt{((4P_{tot} + K_d)^2 - 16P_{tot}{}^2)}}{4}$$

where P_{tot} is the total protein concentration (expressed in terms of monomer) estimated by UV absorbance at 280 nm. The resulting K_d is expressed in terms of the concentration of globin subunits. We observed no higher stoichiometries.

To estimate K_d of the heterodimer–heterotetramer transition in Ancα + Ancβ (or mutant ancestral globins) we performed nMS at variable protein concentrations. Because nMS directly quantifies the abundance of all species in solution, we were able to extract molarities for the α_1/β_1 heterodimer and α_2/β_2 heterotetramers and directly calculate the K_d of their association/dissociation equilibrium, without having to fit a large number of K_d values as part of a coupled set of many equilibria across many homomeric and heteromeric forms. At each concentration, we first calculated the total fraction of subunits that were incorporated into haem-bound heterodimers, including both free heterodimers and heterodimers assembled into heterotetramers, as

$$F_{\alpha\beta} = \frac{2x_{\alpha1\beta1} + 4x_{\alpha2\beta2}}{x_{\alpha1} + x_{\beta1} + 2x_{\alpha1\beta1} + 2x_{\alpha2} + x2_{\beta2} + 4x_{\alpha2\beta2} + 2y_{apo-\alpha1\beta1}}$$

where x is the sum of the signal intensities of all peaks corresponding to the stoichiometry indicated by the subscript. $y_{apo-\alpha1\beta1}$ is the signal intensity of the peaks corresponding to heterodimers that are only partially haem-bound and cannot associate into tetramers. The concentration of all haem-bound subunits incorporated into heterodimers (free heterodimers or assembled into heterotetramers) was calculated as $C_{\alpha\beta} = F_{\alpha\beta} \times P_{tot}$. The fraction of all heterodimers incorporated into heterotetramers was calculated as $F_{\alpha2\beta2} = 4x_{\alpha2\beta2}/(2x_{\alpha1\beta1} + 4x_{\alpha2\beta2})$.

Assembly of heterodimers into heterotetramers as concentration increases was then analysed to find the best-fit value of K_d using nonlinear regression and the following equation:

$$F_{\alpha2\beta2} = \frac{1}{C_{\alpha\beta}} \times \frac{4C_{\alpha\beta} + K_d - \sqrt{\left(4C_{\alpha\beta} + K_d\right)^2 - 16C_{\alpha\beta}{}^2}}{4}$$

The resulting K_d is expressed in terms of the concentration of globin subunits contained in heterodimers and heterotetramers.

For homotetramerization of globins expressed in isolation, the K_d of the dimer–tetramer transition was calculated using a similar approach. The fraction of all subunits incorporated into homodimers (including both free homodimers and those associated into homotetramers) was calculated as $F_d = (2x_d + 4x_t)/(x_m + 2x_d + 4x_t)$, and the concentration of all dimers was calculated as $C_d = F_d \times P_{tot}$. The fraction of all dimers that were incorporated into tetramers was calculated as $F_t = 4x_t/(2x_d + 4x_t)$. Nonlinear regression was then used to fit K_d to the data using the equation

$$F_t = \frac{1}{C_d} \times \frac{4C_d + K_d - \sqrt{(4C_d + K_d)^2 - 16C_d{}^2}}{4}$$

The resulting K_d is expressed in terms of the concentration of globin subunits contained in homodimers and homotetramers. For Fig. 3c, Ancα/β4 was coexpressed with Ancα and fractionated by SEC, and the tetrameric fraction was analysed by nMS.

Native MS spectra for human Hb and Ancα/β14 + Ancα at high concentrations contained peaks corresponding to dimers that had lost one or both haems. In these cases, we calculated K_d values by both including and excluding these species. For the fits shown in Figs. 1d and 3d, these peaks were excluded from the analysis; for the fits shown in Extended Data Fig. 2k, they were included. Both approaches yielded K_d estimates of the same order, although the fit to the data was much better in the former case. Spectra for Ancα + Ancβ included twinned peaks, which represent caesium iodide adducts on tetramers. For the fits shown in Figs. 1c and 3d, these peaks were excluded; for the fits in Extended Data Fig. 2i, they were included. Both approaches gave almost identical K_d estimates, although the fit to the data was better in the former case.

Hydrogen–deuterium exchange mass spectrometry

All chemicals and reagents were purchased from Sigma Aldrich (Gilligham, UK). Native equilibration buffer contained 100 mM PBS (H_2O), pH 7.4. Labelling buffer contained 100 mM PBS (D_2O), pD 7.4. Quench buffer contained 100 mM potassium phosphate (H_2O), pH 1.9, with 1 M guanadinium chloride. Five microlitres of protein sample was diluted into 55 µl of a deuterated buffer of the same composition and corresponding pD. This resulted in a labelling solution ~92% D_2O. Samples were incubated for between 15 s and 1 h at 20 °C before being quenched with an ice-cold H_2O buffer (pH 1.9) of equal volume. The quenched solution pH was ~2.5 at 0 °C. This was quickly injected into an on-line HDX manager (Waters, Milford, MA, USA). The sample was injected onto a 50-µl sample loop at 0 °C before passing over an immobilized pepsin column (Enzymate Pepsin 5 µm, 2.1 mm × 30 mm, Waters) at 20 °C using an isocratic H_2O (0.1% v/v) formic acid solution (200 µl/min). Peptide products were collected on a trapping column (BEH C18, 1.7 µm, 2.1 mm × 5 mm, Waters) held at 0 °C. After 2 min of collection, and de-salting, peptides were eluted from the trap column on to an analytical column (BEH C18, 1.7 µm, 1 mm × 100 mm, Waters) for separation using a reverse-phase gradient with a flow rate of 40 µl/min. The elution profile using a H_2O/MeCN (+0.1% formic acid v/v) gradient was as follows: 1–7 min 97% water to 65% water, 7–8 min 65% water to 5% water, 8–10 min hold at 5% water. The analytical flow rate was 40 µl/min and the eluate was electrosprayed directly into a Synapt G2Si (Waters, Wilmslow, UK) Q-ToF instrument for mass analysis.

Sample handling was semi-automated using a robotic liquid handling HDX system (LEAP Technologies, Ringwood, Australia) to ensure reproducibility in timings. A blank and cleaning injection cycle was performed between each labelling experiment. Mass spectrometry conditions were as follows: capillary 2.8 kV, sample cone 30 V, source offset 30 V, trap activation 4 V, transfer activation 2 V. The source temperature was set to 80 °C and cone gas flow 80 l/h, the desolvation temperature was 150 °C and the desolvation gas flow was 250 l/h. LeuEnk was used as an internal calibrant and acquired every 30 s. For reference, back-exchange was estimated separately using lyophilized samples of angiotensin II. Angiotensin II was dissolved into D_2O (pH 4.0) and left for 48 h. After that, the sample was loaded onto the same robotic and UPLC system and analysed after 2 min of trapping to give a back-exchange of 31.8 ± 0.2%.

Peptides were identified, in the absence of labelling, by data-independent MS/MS analysis (MS^E) of the eluted peptides and subsequent database searching in the Protein Lynx Global server 3.0 software (Waters). Peptide fragments were generated in the trap region through collisions with Ar gas (0.4 ml/min). Peptide identifications were filtered according to fragmentation quality (minimum fragmentation products per amino acid: 0.2), mass accuracy (maximum [MH]+ error: 5 ppm), and reproducibility (peptides identified in all MS^E repeats) before their integration into HDX analysis. HDX-MS data were processed in

DynamX 3.0 software (Waters), and all automated peptide assignments were manually verified, with noisy and overlapping spectra discarded. External python scripts were written to generate and analyse the Woods plots from data outputs of DynamX.

Sample concentration was varied to control the relative populations of monomeric and dimeric species of Ancα/β. After dilution into the labelling buffer, Ancα/β concentrations were 0.67, 2, 15, and 75 μM; to avoid significant sample overloading of the column when using high concentrations of Ancα/β, samples were diluted during quenching to give an injection quantity of ~15 pmol. To ensure back-exchange occurred equally across all diluted samples, the final ratio of H_2O:D_2O after quenching was kept constant at 54:46 and the pH of the quench buffer adjusted to pH 2.5. This allowed all concentrations to be compared without correcting for back-exchange. All automated peptide assignments were manually verified, with noisy and overlapping spectra discarded. After processing, a sequence coverage of 91% was achieved with a redundancy of 5.3.

Statistical comparison of peptides

For each peptide in the dilution experiment, the difference in deuterium uptake between different conditions was normalized by dividing the difference by the absolute uptake in the dimeric condition (75 μM). In Fig. 2c, peptides that incorporated deuterons in the monomeric condition at quantities statistically indistinguishable from zero ($P < 0.01$) were excluded. For peptide locations and alternative normalization methods, see Extended Data Figs. 6, 7. A permutation test was used to determine whether relative deuterium uptake by residues at IF1 (or IF2) was significantly different from that of other residues. To eliminate statistical non-independence arising from the fact that many peptides overlap, we constructed a non-overlapping peptide set by subsampling without replacement from the total set of peptides, requiring that selected peptides do not share any residues. One thousand such non-overlapping peptide sets were constructed, and a *P* value was estimated for each set using the following permutation test. Peptides in the nonoverlapping set were partitioned into those containing residues mapping to IF1 and those containing no IF1 residues; a similar approach was used to test for a difference between peptides containing IF2 residues and those containing none; peptides containing residues contributing to both interfaces were excluded. The mean of the measured relative uptake difference over peptides in each partition was calculated, and the difference between the means of the two partitions was determined. A null distribution was then estimated by randomly partitioning peptides in the nonoverlapping set into two categories (without changing the size of the categories) and calculating the difference in means between the two randomly permuted peptide partitions. The *P* value was calculated as the proportion of random partitions in which the difference between peptide category means was greater than or equal to that of the difference for the empirical categories. Extended Data Fig. 5 displays the distribution of *P* values calculated in this way for 1,000 non-overlapping peptide sets. An interface category was identified as having significantly increased uptake if the mean *P* value from this analysis was <0.05.

Homology models for Ancα/β IF1 and IF2

Structural modelling of the Ancα/β monomer was performed using SWISS-MODEL. A deoxy structure of an Hbα monomer contained in recombinantly expressed human haemoglobin (1A3N) was used as the template. Hbα was used because its sequence similarity to Ancα/β is greater than that of any other extant globin. Furthermore, both Hbα and Ancα/β form homodimers in isolation, unlike Hbβ (which is a mixture of dimers and tetramers at similar concentrations) or myoglobin. EMBO PISA[62] was used to identify sites in 1A3N subunits that buried >50% of their surface area at the interfaces or formed intersubunit hydrogen-bonding or salt bridge contacts at either IF1 or IF2. The HADDOCK 2.2 webserver was used to dock two Ancα/β monomers along an IF1 or an IF2 orientation by specifying the corresponding homologous residues (1a3n). The best scoring docked complex was used for all subsequent analyses and visualizations.

Homology models, interface burial, and contact maps for Ancα + Ancβ and Ancα/β14

Structural modelling was performed using SWISS-MODEL. A deoxy structure of recombinantly expressed human haemoglobin (PDB 1A3N) was used as the template for Ancα + Ancβ and for Ancα/β14 + Ancα. The extant Hbα and Hbβ were used as templates because they have higher sequence identity to to Ancα and Ancα/β14, respectively, than any other globin paralogues. EMBO PISA was used to estimate residue burial at the interfaces and to predict hydrogen bonds across interfaces. Residues were classified as contributing to an interface if their solvent-accessible surface area was reduced by >10% in the assembled form relative to the nonassembled form. Van der waals contacts were identified as pairs of cross-interface atoms with centre-to-centre distances <3.5 Å, using a custom script. PyMOL v4.19 was used to visualize and render protein structures. The similarity between interfaces in the homology model and those in X-ray crystal structures of extant haemoglobins was assessed by aligning the Ancα/β14 + Ancα tetramer to Hb from human (1A3N) and rainbow trout (*Oncorhynchus mykiss* 2R1H) (Extended Data Fig. 10).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Reconstructed ancestral sequences have been deposited in GenBank (IDs MT079112, MT079113, MT079114, MT079115). Alignment and inferred phylogeny, raw mass spectra, oxygen-binding data, and homology model coordinates have been deposited at https://doi.org/10.5061/dryad.w0vt4b8mx. HDX-MS data are available at https://doi.org/10.5287/bodleian:5zRrdMB7E.

Code availability

Scripts for analysis for the HDX permutation analysis and identification of contacts between subunits in modelled structures have been deposited at https://github.com/JoeThorntonLab/Hemoglobin-evolution.

40. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).
41. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
42. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
43. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55**, 539–552 (2006).
44. Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* **60**, 685–699 (2011).
45. Hoffmann, F. G. et al. Evolution of the globin gene family in deuterostomes: lineage-specific patterns of diversification and attrition. *Mol. Biol. Evol.* **29**, 1735–1745 (2012).
46. Hoffman, F. G. & Storz, J. F. The α^D-globin gene originated via duplication of an embryonic α-like globin gene in the ancestor of tetrapod vertebrates. *Mol. Biol. Evol.* **24**, 1982–1990 (2007).
47. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
48. Schwarze, K., Singh, A. & Burmester, T. The full globin repertoire of turtles provides insights into vertebrate globin evolution and functions. *Genome Biol. Evol.* **7**, 1896–1913 (2015).
49. Natarajan, C. et al. Expression and purification of recombinant hemoglobin in *Escherichia coli*. *PLoS ONE* **6**, e20176 (2011).
50. Imai, K. *Allosteric Effects in Haemoglobin* (Cambridge Univ. Press, 1982).
51. Bonaventura, C. & Bonaventura, J. Anionic control of function in vertebrate hemoglobins. *Integr. Comp. Biol.* **20**, 131–138 (1980).
52. Weber, R. E. & Jensen, F. B. Functional adaptations in hemoglobins from ectothermic vertebrates. *Annu. Rev. Physiol.* **50**, 161–179 (1988).

53. Isaacks, R. E. & Harkness, D. R. Erythrocyte organic phosphates and hemoglobin function in birds, reptiles, and fishes. *Integr. Comp. Biol.* **20**, 115–129 (1980).
54. Benesch, R., Benesch, R. E. & Enoki, Y. The interaction of hemoglobin and its subunits with 2,3-diphosphoglycerate. *Proc. Natl Acad. Sci. USA* **61**, 1102–1106 (1968).
55. Imaizumi, K., Imai, K. & Tyuma, I. The linkage between the four-step binding of oxygen and the binding of heterotropic anionic ligands in hemoglobin. *J. Biochem.* **86**, 1829–1840 (1979).
56. Grispo, M. T. et al. Gene duplication and the evolution of hemoglobin isoform differentiation in birds. *J. Biol. Chem.* **287**, 37647–37658 (2012).
57. Richard, V., Dodson, G. G. & Mauguen, Y. Human deoxyhaemoglobin-2,3-diphosphoglycerate complex low-salt structure at 2.5 A resolution. *J. Mol. Biol.* **233**, 270–274 (1993).
58. Arnone, A. X-ray diffraction study of binding of 2,3-diphosphoglycerate to human deoxyhaemoglobin. *Nature* **237**, 146–149 (1972).
59. Arnone, A. P. M. Structure of inositol hexaphosphate–human deoxyhaemoglobin complex. *Nature* **249**, 195–197 (1974).
60. Cong, X. et al. Determining membrane protein-lipid binding thermodynamics using native mass spectrometry. *J. Am. Chem. Soc.* **138**, 4346–4349 (2016).
61. Marty, M. T. et al. Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Anal. Chem.* **87**, 4370–4376 (2015).
62. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).

Acknowledgements We thank C. Natarajan for technical advice and the Hb co-expression plasmid and members of the Thornton laboratory for technical advice and comments on the manuscript. Supported by NIH R01-GM131128 and R01-GM121931 (J.W.T.), NIH R01-HL087216 and NSF OIA-1736249 (J.F.S.), NIH T32-GM007197 (C.R.C.-R.), a Chicago Fellowship (G.K.A.H.), BBSRC BB/L017067/1 and Waters Corp. (J.L.P.B.).

Author contributions A.S.P. identified and developed the model system. A.S.P., G.K.A.H., and J.W.T. coordinated the project, interpreted the data, and led writing of the manuscript. A.S.P. performed and interpreted phylogenetic analyses and biochemical assays. A.V.S. and J.F.S. performed and interpreted oxygen binding assays. Y.L. and A.L. performed and interpreted native mass spectrometry experiments. S.A.C. and J.L.P.B. performed and interpreted HDX experiments. C.R.C.-R. performed and interpreted biochemical assays. All authors contributed to writing the manuscript.

Competing interests The authors declare no competing interests.

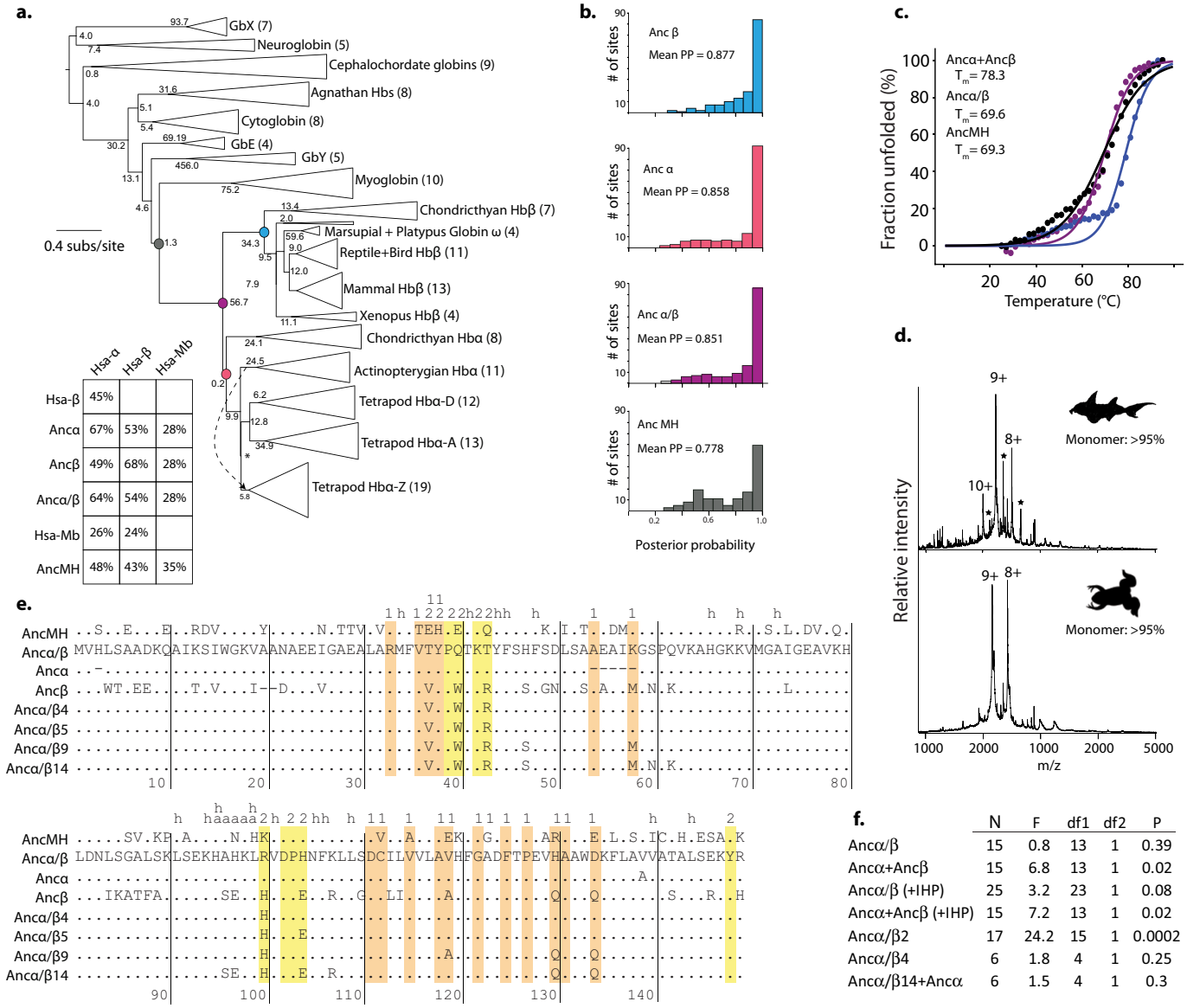
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2292-y.

Correspondence and requests for materials should be addressed to J.W.T.

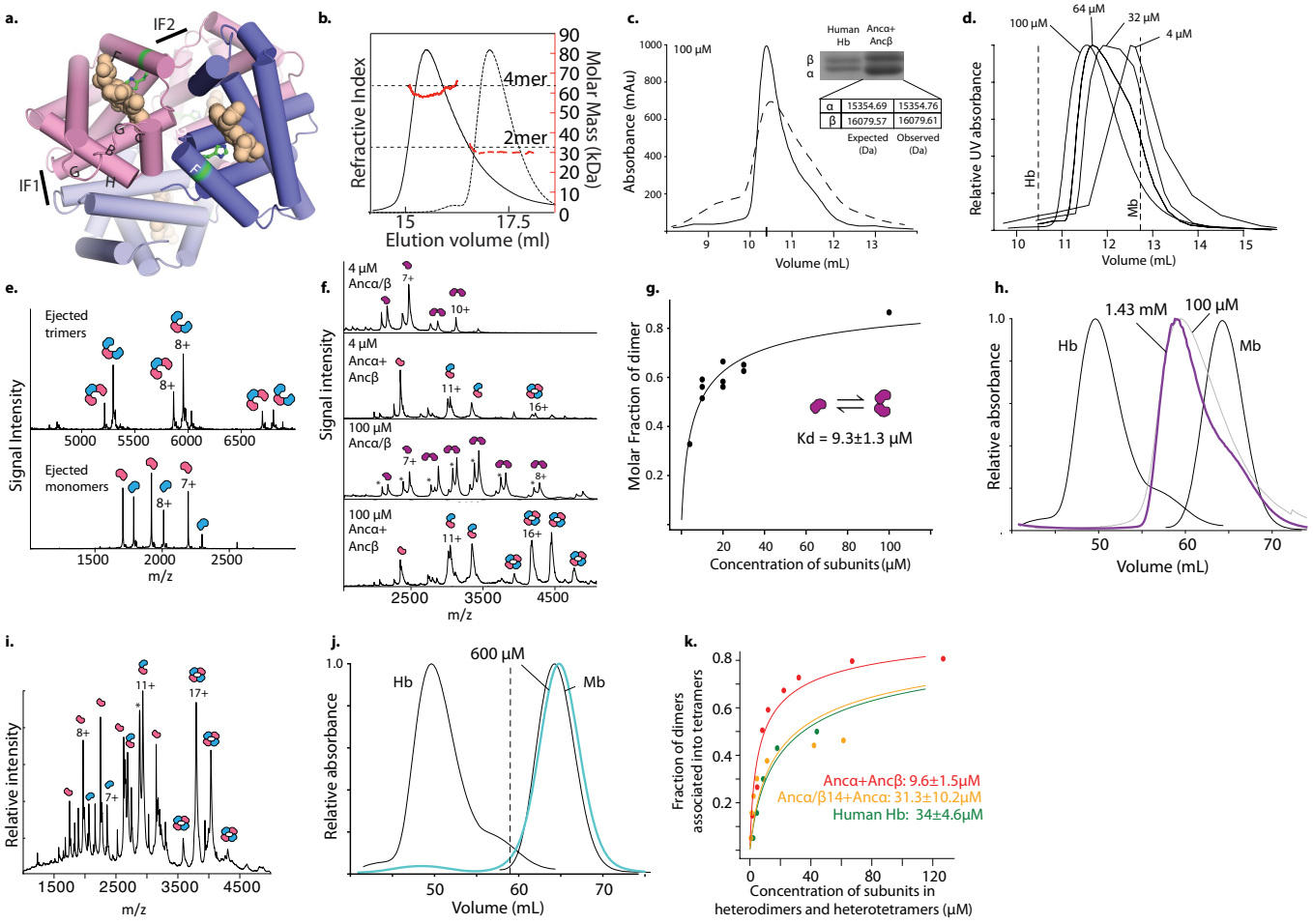
Peer review information *Nature* thanks Nobuhiko Tokuriki and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

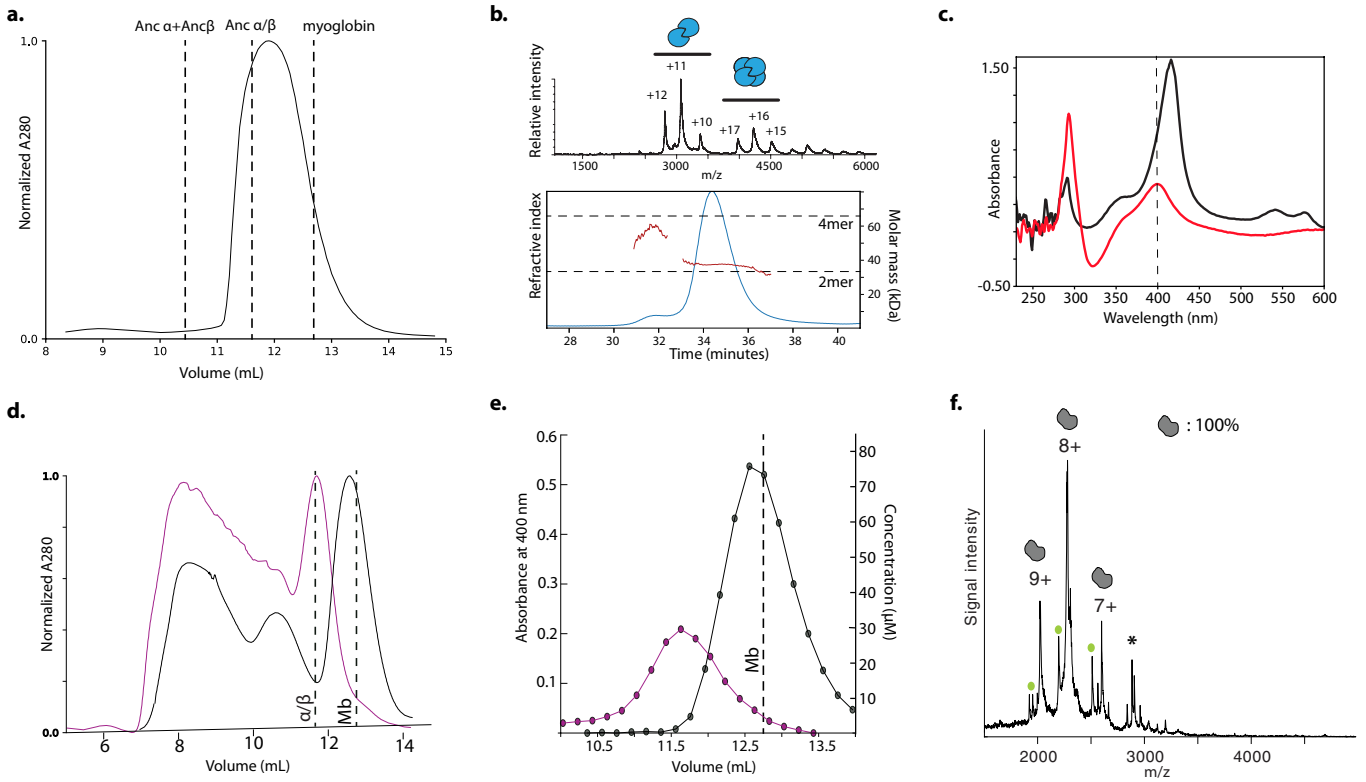


	N	F	df1	df2	P
Ancα/β	15	0.8	13	1	0.39
Ancα+Ancβ	15	6.8	13	1	0.02
Ancα/β (+IHP)	25	3.2	23	1	0.08
Ancα+Ancβ (+IHP)	15	7.2	13	1	0.02
Ancα/β2	17	24.2	15	1	0.0002
Ancα/β4	6	1.8	4	1	0.25
Ancα/β14+Ancα	6	1.5	4	1	0.3

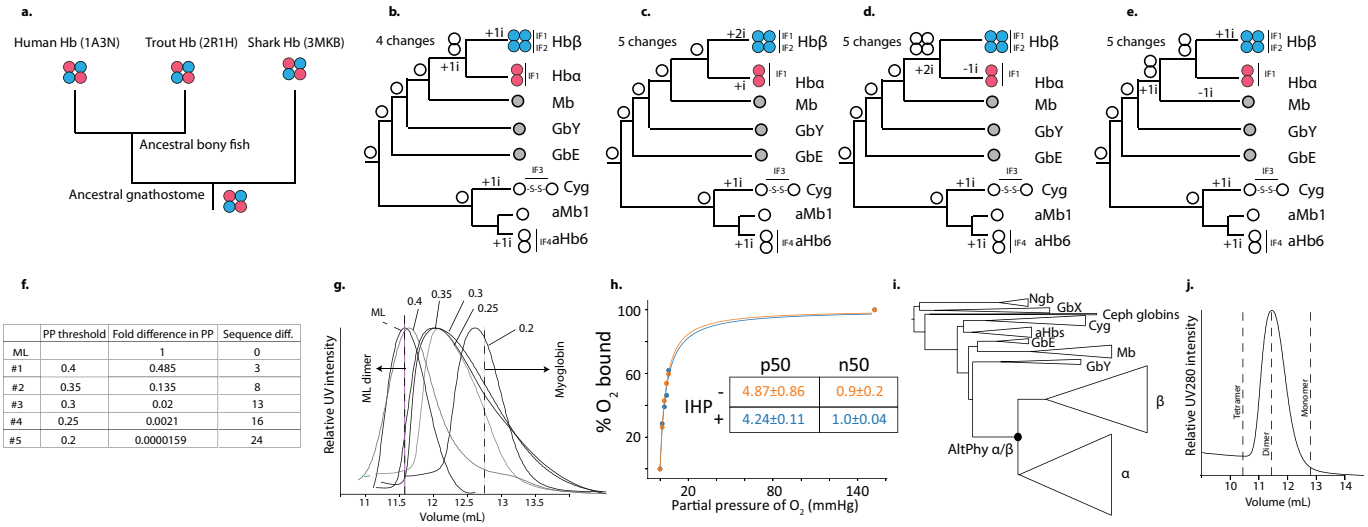
Extended Data Fig. 1 | Reconstruction of ancestral haemoglobin and precursors. a, Phylogeny of Hb and related globins. Node supports are shown as approximate likelihood ratio statistic[59,60]. The number of sequences in each group is shown in parentheses. Ancestral sequences reconstructed in this study are shown as coloured circles. Arrow, branch swap that differentiates this phylogeny from the unconstrained ML phylogeny, which requires additional gene gains and losses. The tree is rooted on neuroglobin and globin X, paralogues that duplicated before the divergence of deuterostomes and protostomes[61]. Inset, pairwise sequence identities among extant (human, Hsa) and reconstructed ancestral globins. **b**, Distribution across sites of the posterior probabilities (PP) of maximum a posteriori states for reconstructed ancestral proteins. **c**, Thermal stability of ancestral globins. Points, fraction of secondary structure lost as temperature increases in Ancα/β (purple), Ancα + Ancβ (blue) and AncMH (black), measured by circular dichroism spectroscopy at 222 nm, relative to signal at 23 °C. T_m and its s.e. were estimated by nonlinear regression; the best-fit curve (lines) are shown. Each point is the mean of four measurements. **d**, Native mass spectra of globin Y from elephant shark (top, *Callorhinchus milii*) and African clawed frog (bottom, *X. laevis*) at 30 μM. Charge states of haem-bound monomer shown. Asterisk, cleavage products. Spectra were collected once. **e**, Sequence alignment of reconstructed ancestral globins. Dots, states identical to Ancα/β; yellow, IF2 sites; orange, IF1 sites; h, sites 4 Å away from the haem; a, sites that link the haem-coordinated proximal histidine (H95) to IF2. **f**, Statistical test of cooperativity of oxygen binding for ancestral proteins and mutants. An *F*-test was used to compare the fit of a model in which the Hill coefficient (n) is a free parameter to a null model with no cooperativity ($n = 1$). Computed *P* value and degrees of freedom (df) are shown. *N*, number of concentrations measured. $^*P < 0.05$. Data were pooled across replicate experiments for nonlinear regression.



Extended Data Fig. 2 | Stoichiometric characterization of ancestral globin complexes. a, Homology model of Ancα + Ancβ (template 1A3N) showing haem (tan spheres). Blue cartoon, Ancβ subunits; red, Ancα. Helices and interfaces are labelled. Green, proximal histidine. **b**, SEC and multiangle light scattering of Ancα/β (90 µM) and Ancα + Ancβ (60 µM). Black, relative refractive index; red, estimated molar mass. Dashed lines, Ancα/β; solid lines, Ancα+Ancβ. Dashed horizontal lines, expected mass for dimers and tetramers. **c**, SEC of human Hb (dashed) and Ancα + Ancβ (solid) at 100 µM. Top inset, SDS–PAGE of these complexes, with bands corresponding to α- and β-subunits. Bottom inset, masses estimated by denaturing MS of Ancα + Ancβ, compared to expected masses based on primary sequence. **d**, SEC of Ancα/β across a series of concentrations. Dashed vertical lines, elution peak volumes of human haemoglobin tetramer and myoglobin monomer. **e**, Tandem MS of the heterotetrameric peak in the Ancα + Ancβ nMS (indicated in Fig. 1b). Ejected monomer and trimer charge series and the subunits they contain are shown. Pink, Ancα; blue, Ancβ. **f**, nMS of Ancα + Ancβ and Ancα/β at 4 µM and 100 µM. Charge series and fitted stoichiometries are indicated. *Unhaemed apo form. **g**, Monomer–dimer association by Ancα/β. Abundances of monomers and dimers were characterized using nMS across a range of concentrations. Circles, fraction of all subunits that were assembled into dimers as a function of the concentration of subunits in all states. Nonlinear regression (line) was used to estimate the dissociation constant (K_d, with s.e.). **h**, SEC of Ancα/β at high concentrations (purple and grey lines). Black curves show SEC traces of human Hb and myoglobin for comparison. **i**, nMS of human Hb at 50 µM. **j**, SEC of AncMH (cyan) at a high concentration. SEC traces of human Hb and myoglobin (black) are shown for reference. Dashed line, Ancα/β dimer elution peak volume (see **f**). **k**, Alternative estimation of affinity of dimer–tetramer association by nMS. For human Hb (green) and Ancα/β14 + Ancα (orange), the fraction of heterodimers incorporated into heterotetramers includes both haem-deficient and holo-heterodimers. For Ancα + Ancβ (red), caesium iodide adduct was included. Compare to Figs. 1d and 3d. K_d values (with s.e.) were estimated by nonlinear regression (lines). All concentrations are expressed in terms of monomer. All nMS and SEC experiments were performed once at each concentration.

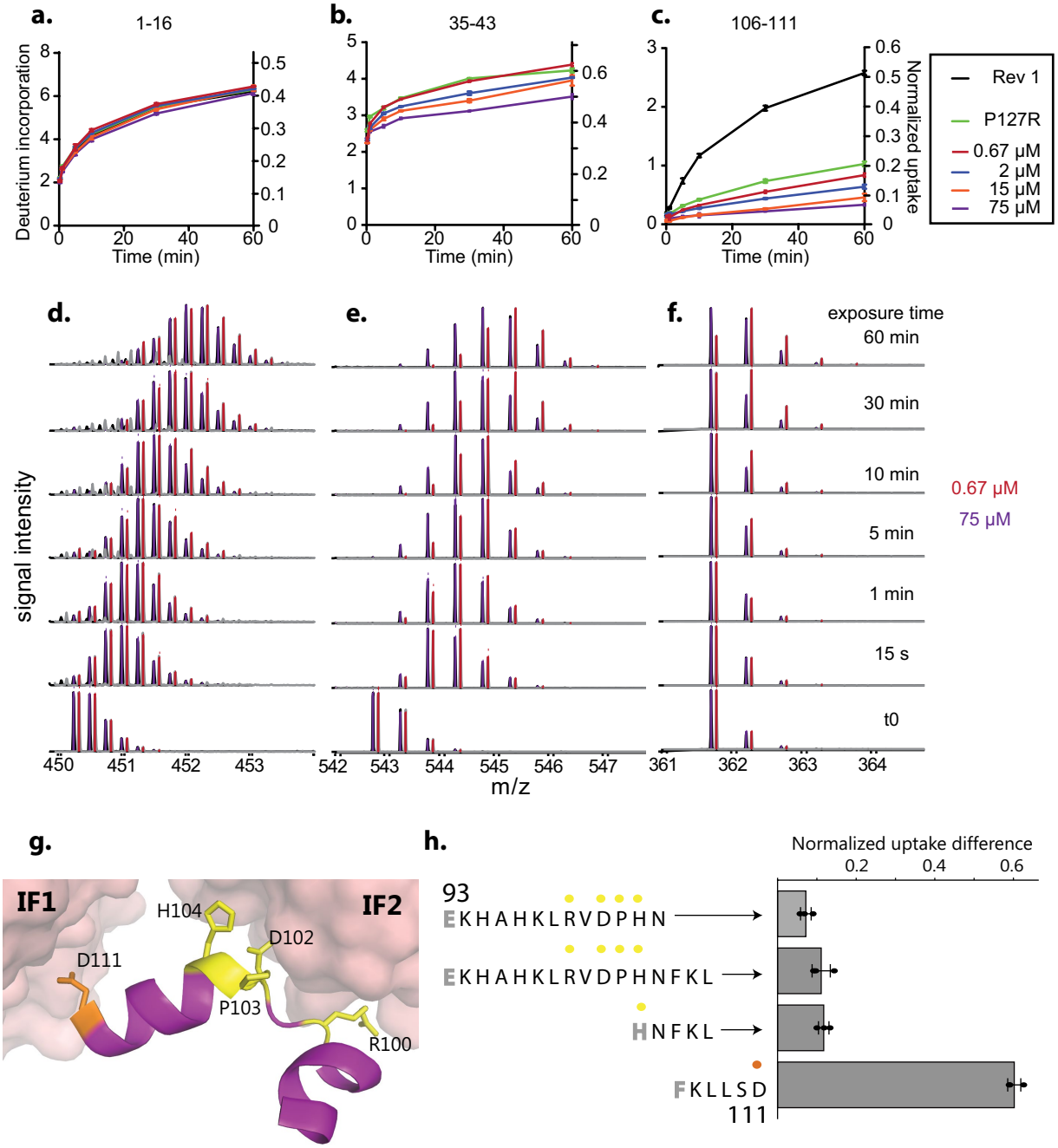


Extended Data Fig. 3 | Stoichiometric analysis of Ancα, Ancβ, and AncMH. **a**, SEC of Ancα at 75 μM. **b**, nMS spectrum (top, at 20 μM) and SEC–MALS (bottom) of Ancβ. Blue, UV absorption; red, molar mass estimated by light scattering. **c**, Colorimetric haemoglobin concentration assay. Absorbance spectra before (black) and after (red) adding 150 μl Triton/NaOH reagent to 50 μl purified Ancα/β. In the presence of reagent, globins absorb at 400 nm. **d**, SEC of crude cell lysate after expression of AncMH (purple) and Ancα/β (black). Dashed lines, expected elution volumes for monomer (human myoglobin) and dimer (Ancα/β). **e**, Colorimetric haemoglobin concentration assay on collected SEC fractions of crude lysate containing AncMH (purple) and Ancα/β (black). **f**, nMS of His-tagged AncMH at 70 μM, with monomer charge series indicated. *Cleavage product. Green, apo. Fractional occupancy of the monomeric form is shown. All experiments were performed once.

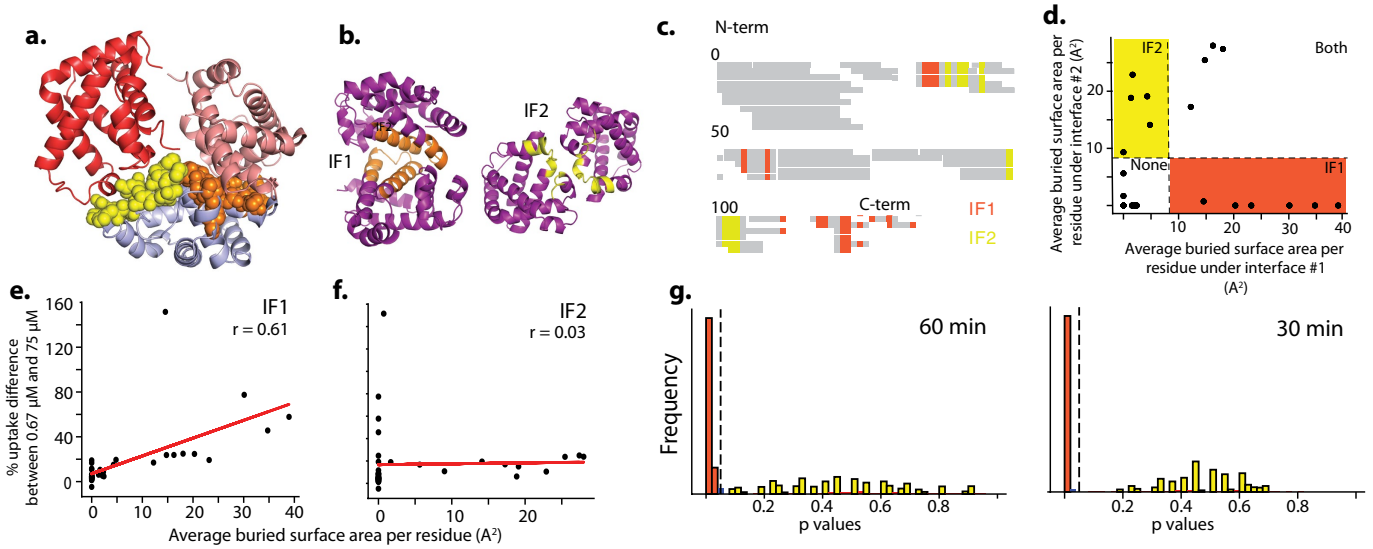


	PP threshold	Fold difference in PP	Sequence diff.
ML		1	0
#1	0.4	0.485	3
#2	0.35	0.135	8
#3	0.3	0.02	13
#4	0.25	0.0021	16
#5	0.2	0.0000159	24

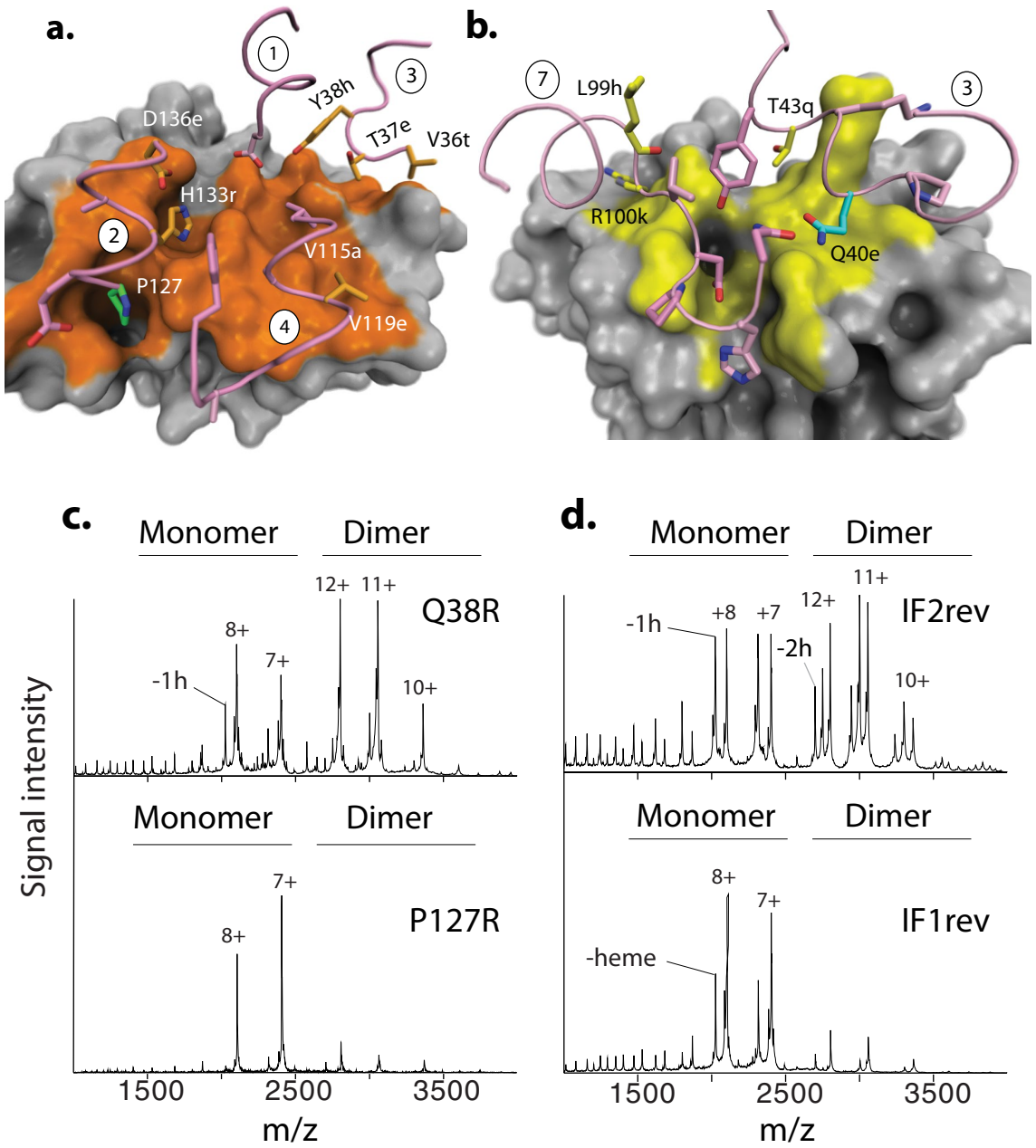
Extended Data Fig. 4 | Biochemical inferences about ancestral Hbs are robust to uncertainty in sequence reconstructions. a–**e**, Maximum parsimony inferences of ancestral stoichiometry and interface losses or gains based on the distribution of stoichiometries among extant globins. **a**, Hbs in all extant lineages of jawed vertebrates are heterotetramers, supporting the inference that AncHb was heterotetrameric. Stoichiometries from representative species' Hbs are shown with PDB IDs. **b**–**e**, Each panel shows a hypothetical set of ancestral stoichiometries, plotted on the phylogeny of extant Hb subunits and closely related globins, with the minimal number of changes required by each scenario. **b**, The most parsimonious reconstruction is that Ancα/β was a homodimer and AncMH was a monomer. **c**, For Ancα/β to have been a tetramer, early gain and subsequent loss of IF2 in Hbα would be required. **d**, For Ancα/β to have been a monomer, IF1 would have to have been independently gained in Hbα and Hbβ. **e**, For AncMH to have been a dimer, IF1 would have to have been lost in lineages leading to the monomers myoglobin (Mb) and globin E (GbE)[12,13]. The dimeric globins most closely related to Hb—agnathan 'haemoglobin' (aHb) and cytoglobin (Cyg)—use interfaces that are structurally distinct from those in Hb[15,16], indicating independent acquisition. **f**–**j**, Alternative reconstructions of Ancα/β are biochemically similar to the ML reconstruction. **f**, Alternative ancestral versions of Ancα/β were constructed, each containing the the ML state at every unambiguously reconstructed site and the second most likely state at all ambiguously reconstructed sites, using different thresholds of ambiguity. For each alternative reconstruction, the table shows the threshold posterior probability (PP) used to define an ambiguous site, as well as the fold-difference in total PP of the entire sequence and the number of sites that differ from the ML reconstruction. **g**, SEC at 75 μM of ML reconstruction of Ancα/β and AltAll reconstructions, which contain all plausible alternative states with PP above a threshold. Dashed lines show elution peak volumes for the dimeric ML α/β and monomeric human myoglobin. Constructs that elute between the expected volumes for dimer and monomer indicate dimers that partially dissociate during the run. None tetramerize; all form predominantly dimers, except AltAll(PP >0.2), which is ~62,000 times less probable than ML, which is mostly monomeric. UV traces were collected once for each construct. **h**, Oxygen binding curves of Ancα/β-AltAll(0.25), the dimeric AltAll with the lowest PP, with and without 2× IHP. Dissociation constant (P50, with s.e.) estimated by nonlinear regression is shown. Lack of cooperativity is indicated by the Hill coefficient (n50 = ~1.0). Oxygen binding at each concentration was measured once. **i**, Alternate globin phylogeny that is more parsimonious than the ML topology with respect to gene duplications and synteny but has a lower likelihood given the sequence data. A version of Ancα/β (Ancα/β-AltPhy) was reconstructed on this phylogeny. **j**, SEC of Ancα/β-AltPhy. Dashed lines show expected elution volumes for various stoichiometric forms.



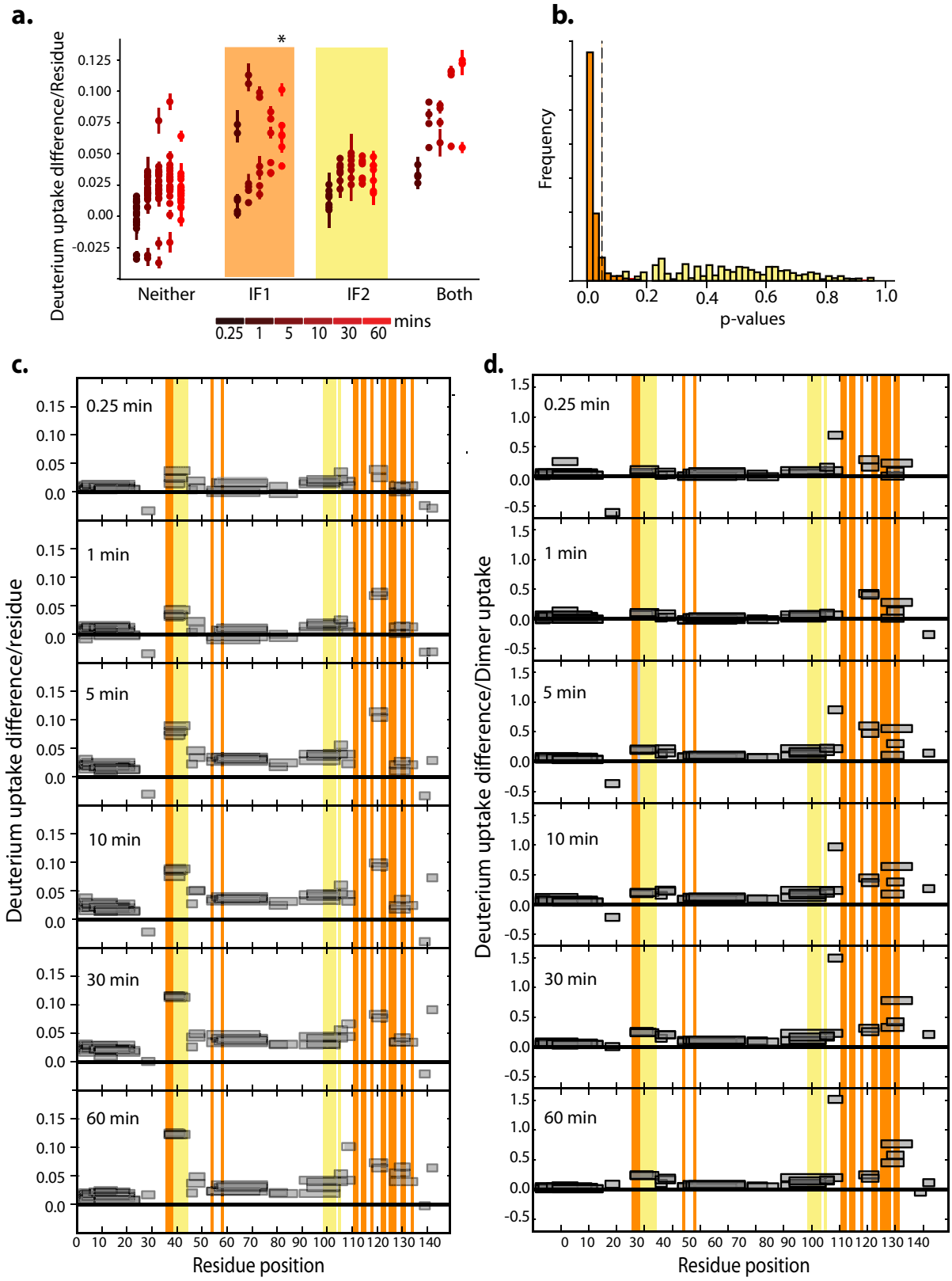
Extended Data Fig. 5 | HDX-MS of Ancα/β. a–**c**, Deuterium uptake measurements across time for three peptides. Left vertical axis, raw deuterium incorporation; right vertical axis, deuterium incorporation divided by the total number of exchangeable amide hydrogens per peptide. Uptake curves for four concentrations of mutants IF1rev and P127R are shown. Each point shows mean ± s.e. of three replicate measurements. **d**–**f**, Raw MS spectra for the peptides shown in **a**–**c**, respectively, at 0.67 μM (red, at which the protein is monomeric), and 75 μM (purple, at which it is entirely dimeric: see Extended Data Fig. 2). The traces are slightly offset to allow visualization. One replicate at each incubation time is shown. **g**, Amino acids 99 to 111 contact IF1 (orange) or IF2 (yellow). The homology model of one chain of Ancα/β (cartoon and sticks) was aligned to the α-subunit of human Hb (PDB 1A3N); β-subunits are shown as surfaces. **h**, Normalized deuterium uptake difference (mean ± s.e. from three replicates), defined as the uptake difference between monomer and dimer divided by the uptake of the monomer, observed for peptides containing amino acids 99–111. Grey N-terminal residues do not contribute to uptake. Amino acid sequences are aligned and labelled (orange dots, IF1; yellow dots, IF2).



Extended Data Fig. 6 | Statistical analysis of HDX-MS results for peptides containing interface residues. a, Residues in human Hb (PDB 1A3N) that bury at least 50% of their surface area in either IF1 (orange) or IF2 (yellow) are shown as spheres. Red and pink, α-subunits; blue, β-subunits. **b**, Homology models of Ancα/β dimer across IF1 (left) and IF2 (right). Two subunits of Ancα/β were computationally docked using HADDOCK using the α1/β1 interface (IF1, left) or α1/β2 interface (IF2, right) of human Hb (1A3N) as a template. **c**, Coverage of peptides produced by trypsinization of Ancα/β, assessed by MS. Orange and yellow, sites that bury surface area at IF1 and IF2 in the modelled dimeric structures, respectively. **d**, Classification of trypsin-produced peptides that contribute to IF1 or IF2. Each circle represents one peptide, plotted by average surface area per residue buried at each interface (total buried area divided by total number of residues). Dashed lines, cutoffs to classify peptides as contributing to IF1 (orange) or IF2 (yellow). **e**, **f**, Correlation between change in deuterium uptake and burial of surface area at IF1 or IF2. Each point is one of 47 peptides, plotted according to the normalized difference in deuterium uptake between concentrations at which monomer or dimer predominates (0.67 or 75 μM, normalized by uptake at 75 μM) and average buried surface area at IF1 or IF2. *r*, Pearson correlation coefficient. **g**, Permutation test to evaluate the difference in deuterium uptake at two time points by peptides containing IF1 versus all other peptides (orange), or IF2 versus all other peptides (yellow). To avoid non-independence, the experimental data were reduced to a set of nonoverlapping peptides by sampling without replacement. Peptides were categorized by whether they contained residues at IF1, IF2, or neither; peptides that contributed to both IFs were excluded. For each interface, the mean uptake by peptides contributing to the interface was calculated, as was the mean uptake by peptides not in that category, and the difference in means was recorded. Peptide assignment to categories was then randomized, and the difference in mean uptake recorded; this permutation process was repeated until all possible randomized assignment schemes for those peptides had been sampled once. *P* value, fraction of permuted assignment schemes with a difference in mean uptake between categories greater than or equal to that from the true scheme. This process was repeated for 1,000 nonoverlapping peptide sets; the histogram shows the frequency of *P* values across these sets. Dashed line, $P = 0.05$.

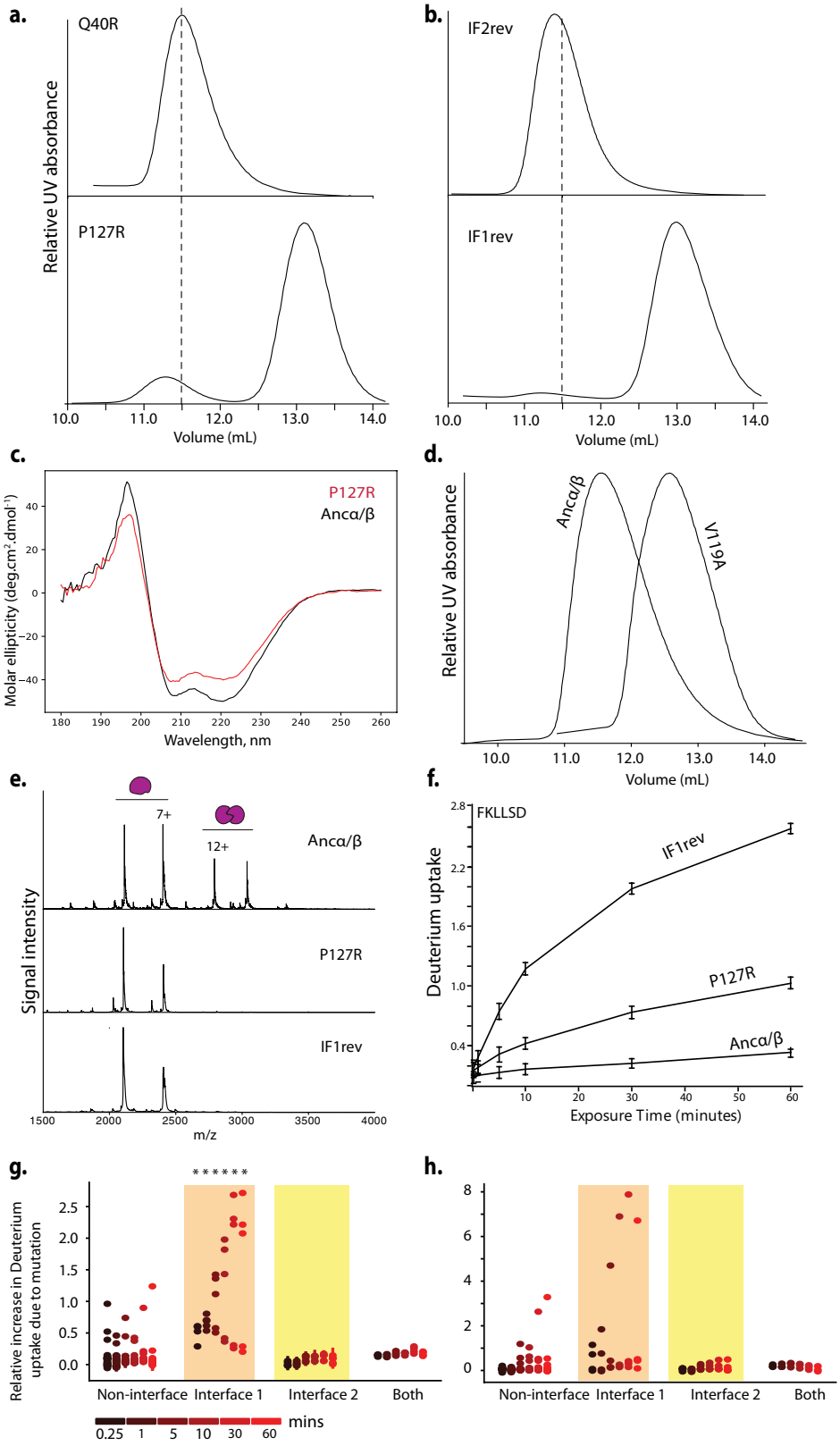


Extended Data Fig. 7 | Dissection of IF1 and IF2 by HDX-MS and mutagenesis. a, **b**, Peptides with residues contributing to IF1 (**a**) or IF2 (**b**) that have the largest relative uptake difference upon dimerization are shown as purple tubes. Sticks, side chains predicted to contact the other subunit (orange surface, IF1; yellow surface, IF2). Side chains are coloured orange (IF1) or yellow (IF2) if they were substituted between AncMH and Ancα/β; purple, unchanged in that interval; green, site for targeted mutation P127; blue, Q40. Circled numbers show the rank of each peptide among all peptides for the normalized difference in deuterium uptake between monomer and dimer conditions. Homology models of the Ancα/β dimer using half-tetramers of human Hb (1A3N) are shown. In **a**, the dimer is modelled using the α1/β1 subunits; in **b**, it is modelled on the α1/β2 subunits. **c**, **d**, nMS of interface mutants Q40R (at IF2) and P127R (at IF1) and for mutants IF1rev and IF2rev, in which interface residues in Ancα/β were reverted to their states in AncMH. All assays at 20 μM. Stoichiometries and charge states are labelled. Unhaemed peak series due to haem ejection during nMS is labelled. Spectra were collected once.



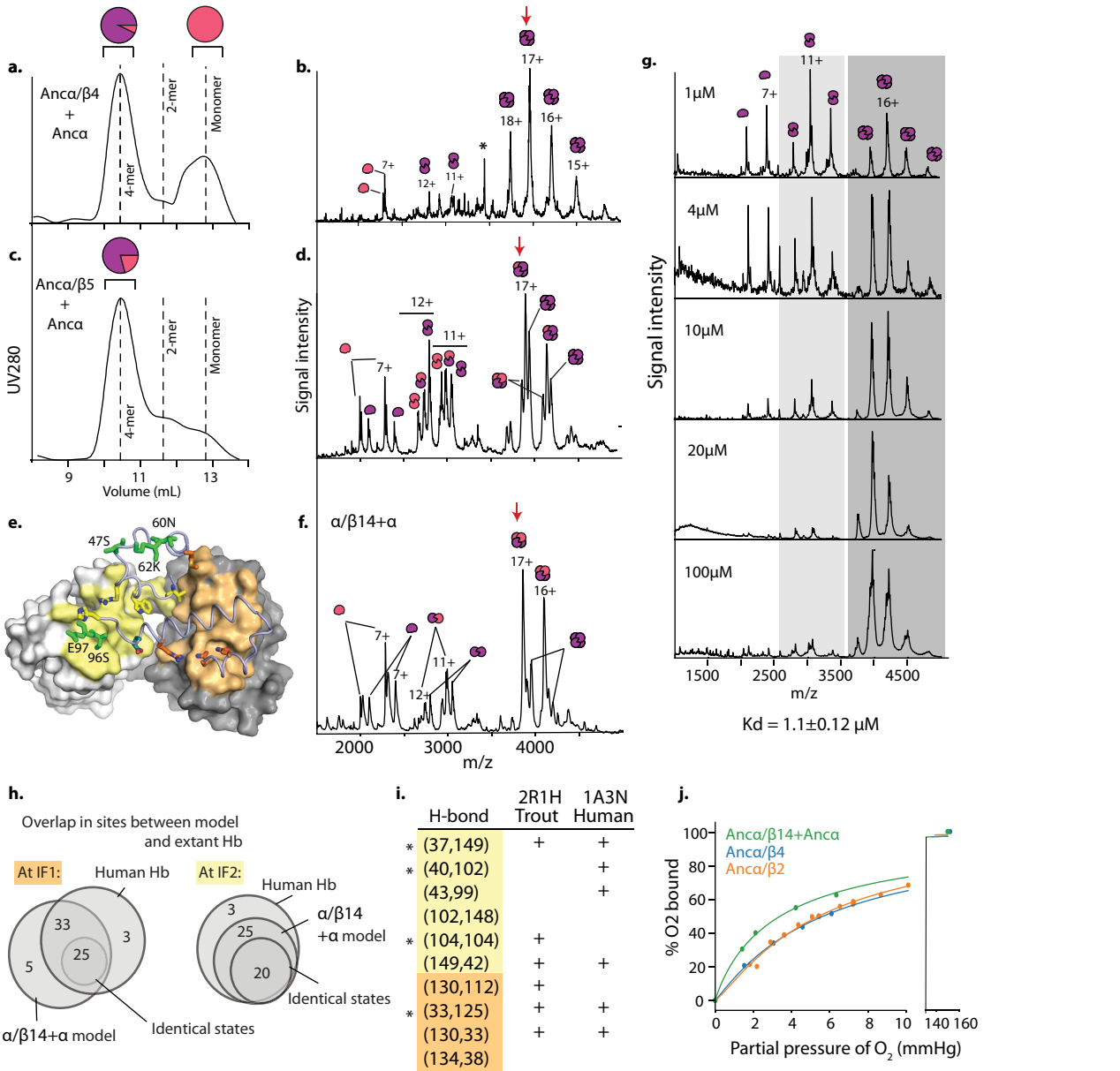
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Alternative methods to normalize deuterium uptake. a, Deuterium uptake difference between monomer (0.67 μM) and dimer (75 μM) at each time point was normalized by the length of each peptide. Peptides were categorized by the interface to which they contribute, as in Fig. 2c. *Interface peptide sets that show significantly increased uptake upon dilution when compared to peptides outside of that interface, as determined by a permutation test (see Extended Data Fig. 6). Each point shows the mean ± s.e. from three replicates. **b**, Permutation test to evaluate the difference in deuterium uptake at 60 min by peptides at each interface, when uptake difference per peptide is normalized by length (as described in Extended Data Fig. 6g). Orange, peptides with IF1-containing residues versus those with no IF1 residues. Yellow, IF2-containing peptides versus those with no IF2 residues. Dashed line, $P = 0.05$. **c**, **d**, Average deuterium uptake difference per residue (**c**) and uptake difference normalized by dimer uptake (**d**) for peptides at different time points. Orange, IF1 sites; yellow, IF2 sites. Each rectangle shows the position of the peptide in the linear sequence and its uptake (mean of three replicates).

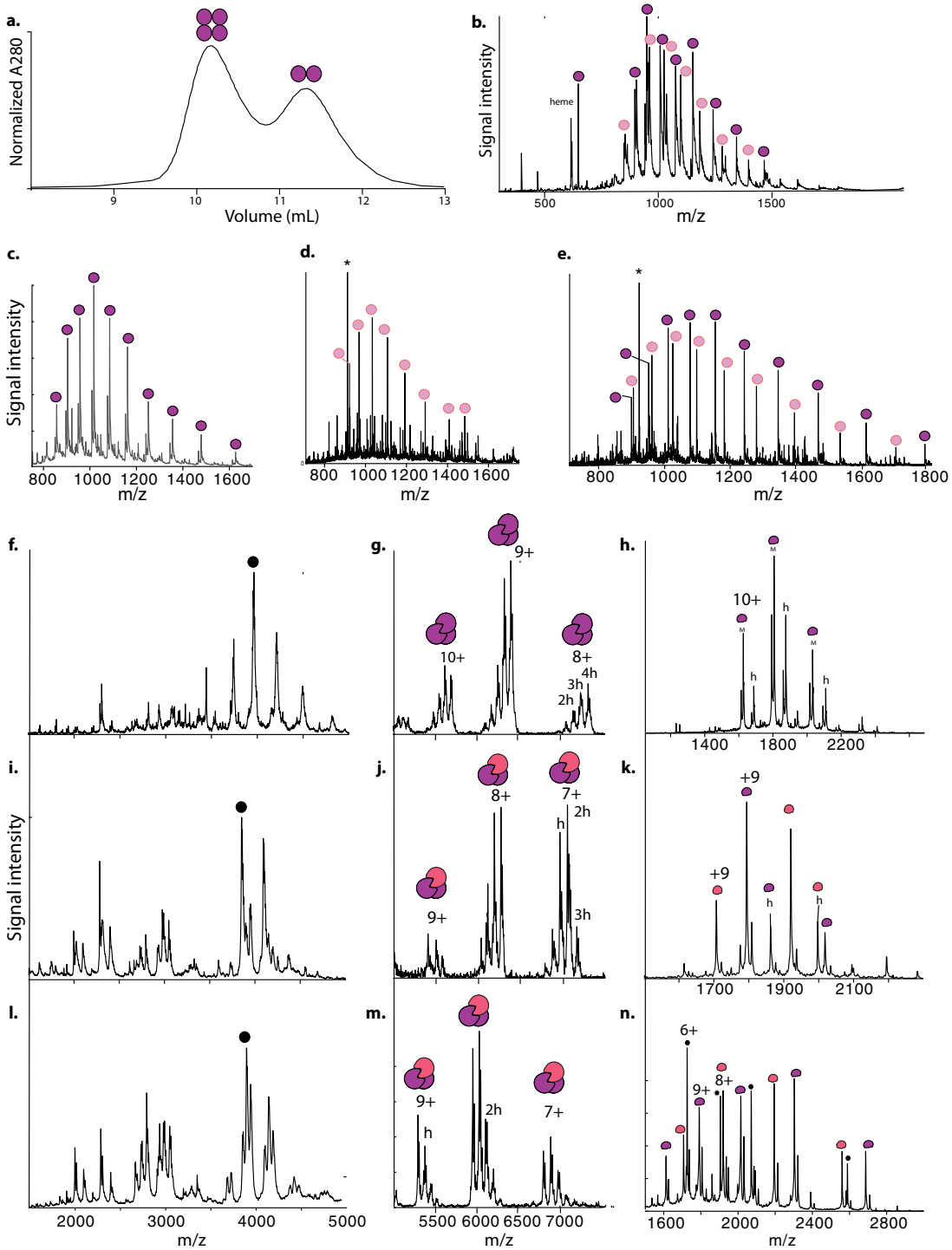


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Effect of interface-disrupting mutations on Ancα/β. **a**, **b**, SEC of mutants at IF2 (Q40R and IF2rev, which reverts all substitutions that occurred between AncMH and Ancα/β at IF2 sites) and at IF1 (P127R and IF1rev) at 100 μM. Dashed line, elution peak volume for Ancα/β. **c**, Circular dichroism spectra for P127R and Ancα/β, showing comparable helical structure. **d**, SEC from IF1 mutant V119A at 64 μM, compared to Ancα/β. **e**, nMS of Ancα/β, P127R and IF1rev at 10 μM. Stoichiometries and charges are shown. For **a**–**d**, nMS and SEC experiments were performed once per concentration. **f**, Normalized deuterium uptake by IF1-containing peptide 106–111 in HDX-MS of Ancα/β (75 μM) and mutants P127R (2 μM) and IF1rev (2 μM). Mean ± s.e. of three replicates. **g**, **h**, Difference between deuterium uptake by each peptide in Ancα/β and uptake by the same peptide in IF1 mutants P127R (**g**) and IF1rev (**h**), both at 2 μM, normalized by uptake in Ancα/β. Peptides are classified by interface category. Mean ± s.e. of three replicates. *Peptide sets that have significantly increased relative uptake (by permutation test, see Extended Data Fig. 6) compared to all other peptides (peptides containing both IF1 and IF2 residues excluded).



Extended Data Fig. 10 | Genetic mechanisms of tetramer evolution. a, **c**, SEC of Ancα/β containing sets of historical substitutions, when coexpressed and purified with Ancα. Dashed lines, elution volumes of known stoichiometries (4-mer, Ancα + Ancβ; 2-mer, Ancα/β; monomer, human myoglobin). Pie charts, relative proportions of α (pink) and α/β mutant (purple) subunits in fractions corresponding to each peak, as determined by high-resolution MS (Extended Data Fig. 11). **b**, nMS of tetrameric fraction in **a** at 20 μM (monomer concentration). *Apparent impurity. Together, **a** and **b** show that tetramers formed by coexpression of Ancα/β4 + Ancα incorporate virtually no α-subunits. Occupancy from this experiment is shown in Fig. 3b. **d**, **f**, nMS of unfractionated purified protein complexes of Ancα/β5 + α and Ancα/β14 + α at 20 μM. Charge series, stoichiometries indicated. Red arrows, peaks isolated for further characterization by tandem MS (Extended Data Fig. 11). **e**, Homology model of Ancα/β14 + α using Human Hb (1A3N) as template. Yellow and cyan sticks, Ancβ-lineage substitutions on IF2; orange sticks, Ancβ substitutions on IF1; yellow surface, αIF2; orange surface, αIF1; green, five β substitutions close to the interfaces included in Ancα/β14 + α. **g**, nMS of Ancα/β2 across concentrations. Charge series and stoichiometries indicated. **h**, Similarity between interfaces in Ancα/β14 + Ancα homology model and X-ray crystal structure of Human Hb. Venn diagrams show sites buried at IF1 and IF2 in one or both structures. Small circle, number of shared interface sites with identical amino acid state. **i**, Hydrogen-bond contacts at interfaces in Ancα/β14 + α homology model are also found in X-ray crystal structures of extant haemoglobins. Residue pairs hydrogen-bonded in Ancα/β14 + α IF2 (yellow) and IF1 (orange) are listed; +also present in crystal structure; *interactions discussed in the main text. PDB identifiers are shown. **j**, Oxygen equilibrium curves of Ancα/β14 + α, Ancα/β4, Ancα/β2. All experiments were performed once per concentration. Lines, best-fit curves by nonlinear regression.



Extended Data Fig. 11 | Stoichiometric characterization of Ancα/β containing historical substitutions. a, SEC of Ancα/β5. Circles show stoichiometry associated with each peak's elution volume. **b**, High-resolution accuracy mass spectrometry (HRA-MS) of Ancα/β5 + α. Purple circles, peaks associated with Ancα/β5; pink, Ancα. **c**, HRA-MS of tetramer-containing SEC fraction of Ancα/β4 + Ancα. **d**, HRA-MS of monomer-containing SEC fraction of Ancα/β4 + Ancα. *922 *m/z* calibration reference standard. **e**, HRA-MS of Ancα/β9 + Ancα. **f**, nMS of tetramer-containing SEC fraction of Ancα/β4 + Ancα (Fig. 3a, b). Black circle, most abundant peak used for tandem MS. **g**, Tandem MS of isolated most-abundant peak in **f**, showing trimer-containing peaks. Charge states and number of haems (h) in the 8+ peak are indicated. **h**, Monomer-containing (M) peaks. **i**–**k**, nMS (**i**) and tandem MS (**j**, **k**) of Ancα/β14 + Ancα (Fig. 3f) as in **f**–**h**. **l**–**n**, nMS and tandem MS of Ancα/β5 + Ancα (Fig. 3c, d) as in **f**–**h**. Black dots in **n** mark charge species produced by cleavage of Ancα/β5. All experiments were performed once.

nature research

Corresponding author(s): Thornton

Last updated by author(s): 18/02/2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	NCBI BLAST was used to collect sequences from NCBI databases.
Data analysis	As described in the methods section, MAFFT v7 was used to build sequence alignments. PhyML 3.1 was used to infer phylogeny from globin alignment. PAML 4.1 was used to infer ancestral sequences using maximum likelihood. PyMOL v1.3 was used to visualize and render protein structures. MassHunter was used to perform mass deconvolution on high resolution accuracy mass spec. data. UNIDEC v 1.0 was used to fit molar masses to and estimate molar abundances from native mass spectrometry. SWISS-MODEL (online server: https://swissmodel.expasy.org/) and EMBO PISA v1.48 were used to model protein structures and identify protein-protein contacts. Custom scripts were used to perform statistical analyses on Hydrogen deuterium exchange data, fit dissociation constants to Native MS data and melting curves to circular dichroism data (see methods). DynamX 3.0 (Waters) was used to process HDX-MS data. OriginPro 2016 was used to fit p50s and Hill coefficient parameters to observed oxygen binding data.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Reconstructed ancestral sequences have been deposited in Genbank (IDs TBA). Homology model coordinates have been deposited in the Protein Model Database (IDs TBA). Alignment and inferred phylogeny and raw mass spectra have been deposited in Dryad (URL TBA). Scripts for analysis for the HDX permutation analysis and identification of contacts between subunits in modeled structures have been deposited at github (https://github.com/JoeThorntonLab/Hb_evolution).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable: experiments were performed on purified stocks of recombinantly expressed proteins. Technical replication of assays is described in the manuscript.
Data exclusions	No data were excluded from the analyses.
Replication	HDX experiments were performed in 3 technical replicates per construct. Measurements of oxygen affinity, cooperativity, and allosteric regulation were performed in 3-5 technical replicates per construct. Native mass spectra and size exclusion chromatography were performed across multiple concentrations with one measurement per construct/concentration, as described in the manuscript. Error associated with replication is reported in the figures and figure legends.
Randomization	Not applicable. The experiments were performed on recombinantly expressed and purified proteins, not on individuals sampled from a population and then assigned to groups.
Blinding	Not applicable. The experiments were performed on recombinantly expressed and purified proteins, not on individuals sampled from a population and then assigned to groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Work

Your story

Send your careers story to: naturecareerseditor@nature.com



DERYA GÜRER

Geologist Derya Gürer in the remote Anatolian Plateau in Turkey.

VOYAGES OF SELF-DISCOVERY

Scientists explain how fieldwork in remote areas prepared them for a pandemic. **By Carrie Arnold**

As many parts of the world continue with or reimpose coronavirus lockdowns, scientists — like everyone else — are feeling the effects of cabin fever. But long stretches of fieldwork in remote areas and months of working at sea have left some researchers more prepared than others. Geologist Derya Gürer at the University of Queensland in Brisbane, Australia, says that her experiences during such trips have taught her how to cope with the circumstances brought about by the pandemic.

She and four other researchers who shared their fieldwork experiences with *Nature* offered the same advice: cut yourself and everyone else some slack. "People might be struggling with something that you are unaware of," says marine biologist Joana Xavier at the University of Porto in Portugal. "It's very important just to be kind."

DERYA GÜRER
CONTROL WHAT YOU CAN

I returned to land in early March after spending 59 days on my first research cruise to map the sea floor in the Southern Ocean, and I've been in self-quarantine ever since. I have also spent time in remote, mountainous regions, usually for several months at a time, as a field geologist.

On this year's voyage, we had one big storm. You're on this huge vessel and you have to trust the people who are guiding it. We were watching a very skilled crew stabilizing the ship. It didn't make the storm go away, but I focused on positive things and on what was within my control. That made me look at my worries in a different way, almost like an outsider. This approach has been helpful while in quarantine.

I started regularly practising yoga on the ship, and, since then, I've done it almost daily. It has helped me to focus and to control my emotions, and to re-evaluate how I am doing every day.

During the entire two months I was on board, I watched only one film. I just kept busy. I've been doing the same thing while in isolation because of the coronavirus. Many of us are still being told to stay in, and we know people are facing hardship and are losing their jobs or their lives, but the one thing that I know I can control is my thoughts.

We have no direct control over what's happening in the world, what governments decide or what rules are imposed. I think it is important to remember that this is not going to be forever. As a scientist, I try to observe what's going on with a bit of curiosity, as well

© 2020 Springer Nature Limited. All rights reserved.

as to observe my own thoughts and emotions without judgement.

Derya Gürer is a geologist at the University of Queensland, Brisbane, Australia.

LAURA OMDAHL
THINK ABOUT OTHERS

I worked for five seasons — from August to February — at McMurdo Station in Antarctica, between 2007 and 2012. I started as the beverage supply clerk and ended up supervising the stores at all three US bases.

At the McMurdo store, we got one shipment per year, usually from a resupply ship that came in January or February, so we had to make everything last for the rest of the year. At McMurdo, all meals are free in the galley. So everything in the store was junk food and fizzy drinks, souvenirs, extra toiletries, that kind of thing.

Doritos are the most sought-after treats down there. Everyone loves Doritos. So I had to put limits on them. Someone would want an entire case and I would say, "I'm sorry, you can have two to three bags a day." People just don't want to think, 'Okay, so if I eat all the Doritos now, I won't have any in January. Or, if someone else eats all the Doritos, I won't have any. Oh, so maybe we should share?' I tried to pick things people liked, but obviously we couldn't get everything. I'd see people when they first got there. They'd be mad that I didn't have their preferred brand of toothpaste.

When you leave the ice, it's amazing. You go to New Zealand, and it has wonderful produce. The first time you go into a market, it feels fantastic, because you're thinking, 'Oh my God, an avocado. Look at that salad. Look at all the fruit, those strawberries.' I still feel that way every time I go into a grocery shop. It hasn't left me — even now, with the shelves stripped of many goods. Why are people complaining? I can get a can of tomatoes. It might not be the exact can of tomatoes I was looking for, but there are some, and we have tons of food to eat.

Laura Omdahl is a former store supervisor at the US Antarctic Program, Antarctica.

CHRIS TURNEY
KEEP A POSITIVE OUTLOOK

It's important not to beat yourself up if you don't achieve as much as you'd like. As part of my research group's work on developing highly detailed records of past environmental conditions preserved in tree rings, ice cores and sequences of peat and lake sediment, I've spent time in the Antarctic. We talk about the A-factor, which is short for Antarctic factor.

If anything can go wrong, it will. COVID-19 is the extreme version of the A-factor. And we all have families and other commitments outside work. So accept that not everything you have planned will happen. And if you can't work as much as you'd like to, that's all right. Almost everyone in the research community is in the same position and it's more important to keep a positive mindset and look out for everyone else than to disappear into a frenzy of work.

On a research field trip, keeping focused on getting everyone home safely is a powerful motivator. And this means looking to the future. A positive outlook is crucial in this regard. I've seen people reflect too much on their predicament, which distracts them from finding a solution. You don't want people withdrawing into themselves, especially if it affects anyone else. Keep regular communication open and make sure team members are working together for the future. Even if you can't find a solution, preparing for whatever comes next and beyond helps people's mental health enormously.

We're social animals and need contact. Social distancing might mean we have to keep a physical distance from one another, but we can use technology to keep the lines of communication open with family, friends and colleagues. Mix it up and you'll stay sane and achieve more in the long term.

Chris Turney is a geoscientist and explorer at the University of New South Wales, Sydney, Australia.

JOANA XAVIER
MAINTAIN A ROUTINE

We've been in quarantine since early March, and my partner and I think that it's important to keep a routine. I learnt about the importance of doing so when I started joining research cruises in 2010 as part of my work on marine sponges. The cruises involved relatively large vessels travelling across the North Atlantic, in the Nordic seas and across the Arctic mid-ocean ridge.

On board, we worked in shifts, usually 12 hours on, 12 hours off. We always had a routine. Not only did that help us to work efficiently in a small, confined space, but it also helped us to unwind when our work was finished. Whether it was a film night or a quiet cup of coffee first thing in the morning, the rituals kept us anchored while we were at sea. When we began to self-isolate after coronavirus struck, I used the same strategy without even thinking about it.

From the first week, because we realized the lockdown was going to last much longer than we originally expected, we created a schedule with our children, who are five and seven. We agreed that the routine needed to include everything from playtime to study time, arts and crafts, and home tasks such as cooking and cleaning up. We asked the children if they thought the timetable would work for them and we put everyone's schedule up on the living-room wall. We try to stick to it as much as possible, because this also allows me to work from home.

It is working quite well. We make sure we get enough rest — our moods depend a lot on that. Managing our energy is important. I learnt this on board and it's true now at home: we need our resting time and our fun time.

Joana Xavier is a marine biologist at the University of Porto, Portugal.

RACHEL DOWNEY
STEP AWAY FROM WORK

I did my first research cruise as a physical geographer in 2013 as part of the British Antarctic Survey, and I've done three more since then. The longest was for seven weeks.

One of the most important things I've learnt is that you need to take the time to understand your needs — both work and personal. I have to know when to put the lid on, when to stop working. On some ships, we would finish our shift and all have a drink, just to separate the day. The ship was our place of work and we never got to leave it, but the ritual marked the end of work and then we could play darts or games. We had things like film nights to look forward to. We found ways to make the ship work for work and for play.

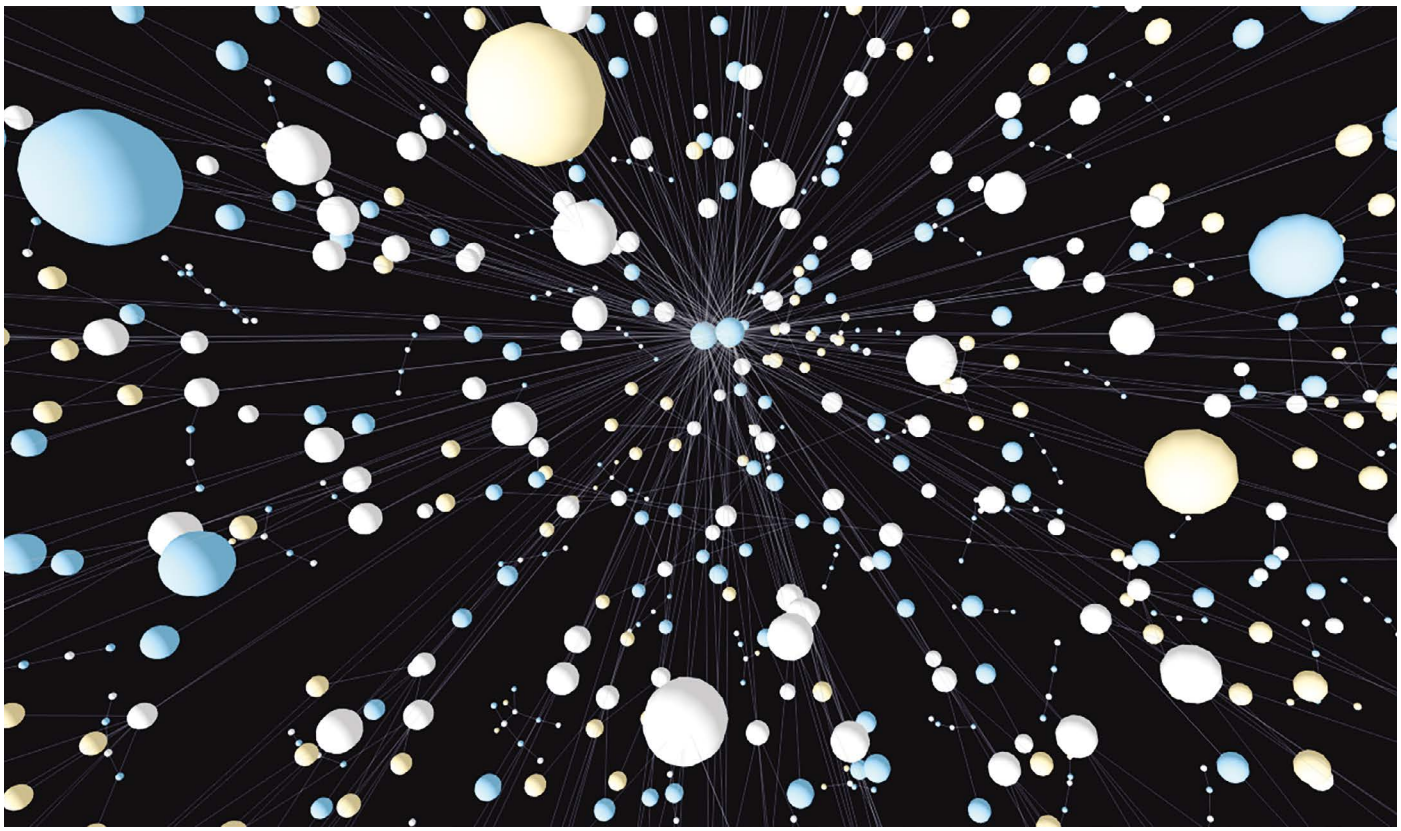
Now that I'm at home, I'm really thankful that we live next to a bush reserve and that we have a garden. It's so nice to get out there after work, with all the soil and the plants. My partner is learning to play the guitar through YouTube. It's so easy for both of us to just keep working, but I've learnt to make myself stop. I try to keep an office space for work in the dining room and then walk away.

This is an unprecedented time, and it's easy to be hard on each other and ourselves and say, 'Oh, I must work twice as hard, or so many hours, or I must achieve this.' You see all these memes about all the incredible things scientists did during the Black Death. I think it's a good time for us to sit back and reflect on our work, and to look at where we are and what direction we want to go in.

Rachel Downey is a marine biologist at the Australian National University, Canberra.

Interviews by Carrie Arnold. Interviews have been edited for length and clarity.

© 2020 Springer Nature Limited. All rights reserved.



J. ROBINSON *ET AL. SCI. SIGNAL.* **13**, EAAZ1482; 2020 (CC BY SA 4.0)

A genome-scale metabolic model of yeast. Each coloured sphere represents a substance that the organism uses in metabolism.

THE MICROBIOME MODELLERS

Computational approaches can reconstruct the complex interactions between gut bacteria and their human hosts. **By Michael Eisenstein**

There is something comforting in the elegance of a chemical reaction. Inputs and conditions on one side of the reaction predictably yield a defined set of products on the other. But this predictability is quickly lost in complex biological systems, where thousands of reactions occur in parallel among vast numbers of interacting cells.

Consider the human gut microbiome, in which roughly 1,000 bacterial species compete and collaborate while communicating with their host. This crosstalk rapidly becomes too complex to capture in a diagram. But understanding these communities is crucial, because their biological activity directly affects our health and susceptibility to disease.

Now, systems biologists are building models to illuminate these black boxes. "The old-school way of looking at hundreds of thousands of reactions is just not feasible, and not desirable either," says Ines Thiele, a microbiome researcher at the National University of Ireland in Galway. "But we have techniques now that try to help us see what's happening much quicker and tackle the emergent complexity that arises."

Thiele and her colleagues are developing mathematical and statistical approaches that use a range of data, from areas such as genomics and biochemistry, to reconstruct highly diverse gut microbial communities computationally. They capture only a fraction of the complex biological reality of the microbiome, but can reveal interactions between microbes and their hosts that would be near-impossible to detect otherwise.

Polished GEMs

For many modelling efforts, the starting point is a genome-scale metabolic model (GEM). Constructed by scanning an organism's genome to determine the biochemical processes it can perform, GEMs are essentially blueprints of the enzymatic assembly lines in every microbe. "They're capturing all of the metabolic capabilities that the cell has," says Jens Nielsen, a systems biologist at Chalmers University of Technology in Gothenburg, Sweden. Researchers can then mathematically model how these inputs and outputs feed into one another.

This approach is particularly powerful for studying the gut, the microbes of which are often difficult to cultivate in the laboratory. Researchers can also draw on existing biological knowledge to extrapolate the possible function of a gene using sequence similarities with known enzymes, for instance. Even sparse information can be useful. Systems microbiologist Karsten Zengler at the University of California, San Diego, and his colleagues, for example, developed a GEM for a relatively under-studied species of marine diatom (a type of plankton), even though functions had been identified for only about one-tenth of its genome[1]. "Surprisingly, it worked," Zengler says. "We have enough information about metabolism to understand what they're doing." His team's modelling approach assigned functions to more than 1,000 diatom genes, which collectively perform nearly 4,500 interconnected biochemical reactions, and predicted where in the cells these reactions are likely to occur.

© 2020 Springer Nature Limited. All rights reserved.

Currently, most microbiome samples are studied by isolating total DNA from a microbial community and breaking it into small pieces for sequencing, a strategy called shotgun metagenomics. The resulting data sets can provide an inventory of the species present in a sample without needing to culture or isolate them. Researchers can combine GEMs to flesh out larger microbial community models, and identify patterns of enzymatic activity on the basis of genes uncovered in the data. They can then thread these processes together to understand which chemicals the particular system is taking up, which products are being released, and which cells are interacting to drive these processes. "You get non-obvious relationships between the microbes that you would just not see by looking at genome sequences themselves," says Thiele.

In one study[2], Zengler's group modelled the dynamic interplay between the alga *Chlorella vulgaris* and the yeast *Saccharomyces cerevisiae*, uncovering unexpectedly high levels of exchange between the two species for certain amino acids. Experimental profiling of metabolites in the sample might have overlooked this exchange, says Zengler, who offers the analogy of trying to infer children's snack preferences by what's left on the table after a birthday party. "You might observe the table and say, 'Look at all those apples – that's what children must like to eat.' But that is not the case."

Not all interactions can be modelled: it's currently impractical to capture the full interplay between all reactions from every cell at the same time. As a result, these models are often best suited for testing existing hypotheses, says Joao Xavier, a systems biologist at the Memorial Sloan Kettering Cancer Center in New York City. "For example, we have several hypotheses about the role of fermentative bacteria that produce short-chain fatty-acid compounds, such as butyrate, that are associated with intestinal health," he says. "So we could try to look for genes that belong to those pathways." Xavier's approach draws on principles used in conventional ecological modelling – for example, equations used for studying predator–prey interactions that capture how changes in one population can affect another. His group then uses machine-learning techniques to model these effects in more complex microbial systems.

But even seemingly simple interactions can hide complexity. Consider the straightforward example of one microbe producing a carbohydrate that another microbe likes to eat. Changes in both the composition of a community and its environmental conditions can influence interactions between species, Zengler notes, citing his yeast–algae study. "In one condition, they loved each other and were best buddies and exchanged everything and grew happily together – better than they did by themselves," he says. But something as simple as changing ammonia levels could make them more adversarial. "This relationship just got worse and worse, and at one point they ended up having a 'divorce' and fighting over their belongings."

"Even seemingly simple interactions can hide complexity."

Just part of the picture

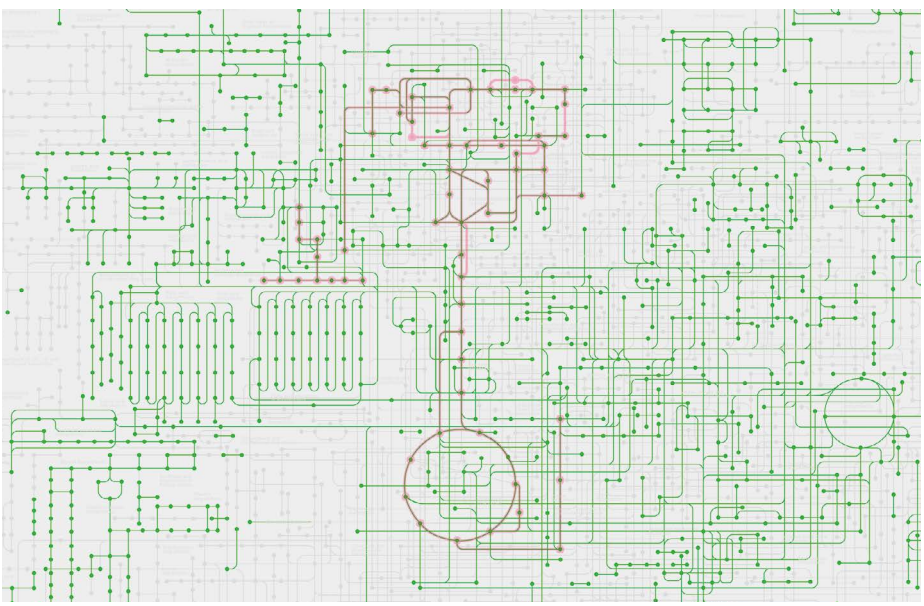
Models can reliably draw such inferences only if they have good underlying data. But those can be difficult to get. The gene databases that researchers use to deduce enzymatic function and other biological activities are still incomplete, and even well-characterized species can hold surprises. In an effort to improve his group's metabolic models, Nielsen tested how well *S. cerevisiae* grew on different carbon-based nutrients[3]. "We found it used many carbon sources that we were not taking into account even in our most state-of-the-art models," he says.

Models of the microbiome in the human gut might also overlook crucial aspects of the host environment, such as the physical structure of the large intestine – a lengthy organ with a bacterial composition that varies from one section to the next. "The geometry of the gut is quite complicated", as are the mechanics of faecal matter travelling through it, says Xavier. There is also extensive communication between the host and microbiome, both at the intestinal barrier and through chemical signals that the intestinal cells subsequently relay throughout the body.

Researchers can indirectly measure the impact of such interactions using samples of blood, urine or cerebrospinal fluid. Thiele has been developing a 'virtual metabolic human' that can integrate these and other data with microbiome models to create a more holistic picture of these interactions in various disease states. "We now have about 30 organs or tissues, arranged in an anatomically accurate manner," she says. Her group is using those models to search for microbial perturbations that might contribute to disorders such as Parkinson's disease.

Xavier's group is collaborating with clinicians to understand how antibiotics affect the gut microbiomes of bone-marrow-transplant recipients, whose immune systems are suppressed to minimize rejection of the transplant. They found that the resulting microbial disruptions can markedly affect the function of the immune system after the transplant, and that supplementing individuals with healthy microbial populations could improve their recovery[4].

Elhanan Borenstein, a systems biologist at Tel Aviv University in Israel, says, "People may underestimate the importance of modelling as a tool. It's not just to simulate a real environment, but also a way to understand first principles." Armed with these fundamentals, researchers might ultimately be able to identify targeted microbiome interventions that meaningfully change clinical outcomes, even if much of the system remains a black box.



M. KANEHISA & S. GOTO *NUCLEIC ACIDS RES.* **28**, 27–30 (2000)

A map of the complex pathways in human metabolism (carbohydrate pathway in red).

Michael Eisenstein is a freelance writer based in Philadelphia, Pennsylvania.

1. Levering, J. *et al. PLoS ONE* **11**, e0155038 (2016).
2. Zuñiga, C. *et al. Nature Microbiol.* **4**, 2184–2191 (2019).
3. Lu, H. *et al. Nature Commun.* **10**, 3586 (2019).
4. Schluter, J. *et al.* Preprint at bioRxiv https://doi.org/10.1101/618256 (2019).

© 2020 Springer Nature Limited. All rights reserved.



Where I work
Andrew Hosie

When you're collecting sea-floor creatures in the abyss off the coast of Western Australia, just about every find is worth a closer look. This beautiful crustacean is a squat lobster — most likely *Galacantha rostrata* is my best guess. A remotely operated vehicle (ROV) spotted it on a bare rock 2.5 kilometres deep in the Ningaloo Canyon System and sucked the creature up using its slurp gun.

I was one of about a dozen scientists aboard the research vessel *Falkor* for a month-long biodiversity survey that was upended by the coronavirus. We left in early March, and were supposed to make a port call in Exmouth halfway through to swap researchers, but that plan got scrapped because of the pandemic. The original crew finished the expedition and collected 30 new species — including swimming worms and heavily armoured barnacles — from depths exceeding 4 kilometres.

The squat lobster didn't survive the trip to the surface, so before it starts degrading, I'm carefully photographing its features with an inexpensive Olympus camera. Sophisticated photography equipment can really eat into a research budget, and this workhorse does the job. Many of the photos will be displayed at the newly renovated Western Australian Museum in Perth, where I'm curator of crustaceans and worms, when it reopens.

We saw some wildly unexpected things on this expedition. There was a metre-tall hydrozoan, related to jellyfish, that stood like a giant flower above the ocean floor. And as the ROV returned to the surface one evening, it passed by a 45-metre-long siphonophore, a string-like colonial organism that is possibly the longest creature ever recorded. The scientists were eating dinner in the galley when the animal appeared on the live video stream. It was an amazing and totally unplanned encounter. Down there, you never know what you'll find next.

Andrew Hosie is curator at the Western Australian Museum in Perth, Australia, and a zoology PhD student at Curtin University in Perth. **Interview by Chris Woolston**.

Photographed by Alex Ingle/SOI.

© 2020 Springer Nature Limited. All rights reserved.

nature

index

South Korea



MISSION RELAUNCH

Drive to spark research creativity

Grand design
An innovation blueprint

Talent magnet
The benefits of mobility

The tables
Institutions leading the way

nature
index

South Korea

Editorial Catherine Armitage, Bec Crew, Gemma Conroy, David Payne, Rebecca Dargie **Analysis** Bo Wu, Catherine Cheung **Art & design** Tanner Maxwell, Annthea Lewis, Wojtek Urbanek **Production** Jason Rayment, Ian Pope, Nick Bruni, Bob Edenbach, Joern Ishikawa **Marketing & PR** Claire Hodge **Sales & partner content** Soon Kim, Lisa Truong **Publishing** Rebecca Jones, Richard Hughes, David Swinbanks.

Nature Index 2020 South Korea, a supplement to *Nature*, is produced by Nature Research, the flagship science portfolio of Springer Nature. This publication is based on data from the Nature Index, a Nature Research database, with a website maintained and made freely available at natureindex.com.

Nature editorial offices
The Campus, 4 Crinan Street, London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

Customer services
To advertise with the Nature Index, please visit natureindex.com or email clientservicesfeedback@nature.com
Copyright © 2020 Springer Nature Limited, part of Springer Nature.

All rights reserved.

Of the top 10 countries in the Nature Index, South Korea is holding its footing better than most amid the tsunami of research from China, despite its rank falling from 8th to 9th place in 2016 – a position it retains.

First, some terms: Share, our key metric, measures contributions to articles in the 82 selected natural-sciences journals in the Index, based on the proportional contribution to each article of authors by country (or institution). Adjusted Share accounts for small changes in the number of Index-journal articles over time.

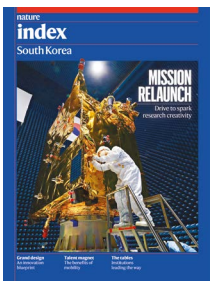
Now to South Korea's performance: China's gain of 63.5% in adjusted Share in the Index from 2015 to 2019 translates to declines for other countries. It's a zero-sum equation: the Index journals publish a finite number of articles. But among the top 10 countries other than China, South Korea, with a mere 6.4% decline in adjusted Share over the past four years, fares better than all but Switzerland (+2.1%) and Australia (−3.8%).

Some of South Korea's research resilience may be due to the fact that it achieved much greater growth in collaboration with China than it did with any other top 10 country. Its bilateral Collaboration Score with China (the sum of Shares on articles with authors from both countries) has increased by 140% since 2015. In 2018, China displaced Japan as South Korea's second-most collaborative partner in the Index, after the United States.

South Korean science is poised for further change. Increased competition from China, plus the fallout from the COVID-19 pandemic, pose challenges. Many hope these will be addressed by a new deal for science and technology comparable to President Moon's Green New Deal re-election manifesto, which promised to reduce South Korea's net carbon emissions to zero by 2050.

This year, as the first centres of the government's flagship Institute for Basic Science come up for review, decisions about which centres to keep, close or open will further point to how South Korea plans to forge its own scientific path. It's a long road between making a pledge and achieving an ambition, but the creativity and determination it has demonstrated by its post-war transformation into a global leader in innovation, as detailed in this supplement, augur well for further success.

Catherine Armitage
Chief editor



On the cover
The Geostationary Korea Multi-Purpose Satellite 2B (Cheollian 2B) continuously observes environmental and marine conditions around the Korean Peninsula. Cover image: Korea Aerospace Research Institute

Contents

S50 Starting small to spark creativity
South Korea's flowing research funds are being redirected to promote original discovery.

S53 National vision
The vast Institute of Basic Science network may be adjusted after a planned review.

S54 A top-down reinvention
Strong investment and systemic reform have cemented South Korea's role as an innovation leader.

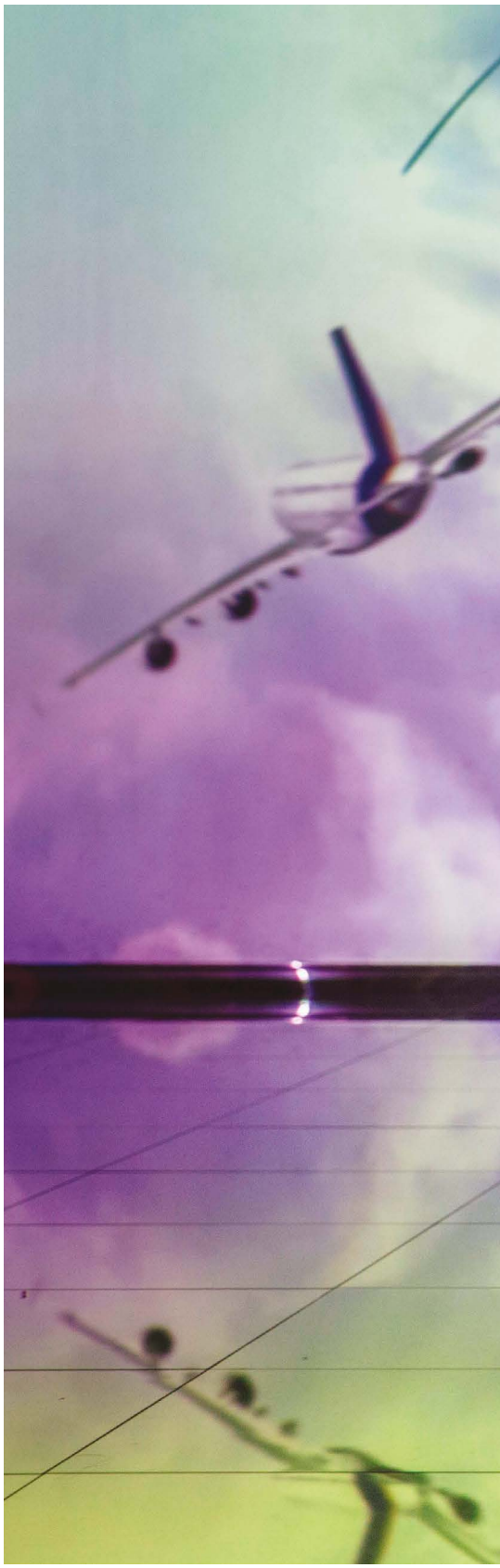
S66 Upwardly mobile
With a greater influx of researchers from abroad, the country's research landscape is more productive.

S72 The tables
The top 50 institutions ranked by article share

© 2020 Springer Nature Limited. All rights reserved.

Starting small to spark creativity

South Korea's flowing research funds are being redirected to basic-research grants to raise competitiveness by encouraging original discoveries. **By Mark Zastrow**



When it comes to funding research, South Korea is in rarefied air. As a percentage of its gross domestic product, its spending on research and development is second only to Israel's.

The speed of this ascent has been dizzying, up from 2.1% of its gross domestic product in 2000, when it was equal to the average of Organisation for Economic Co-operation and Development member nations, to more than 4.5% in 2018, with Israel alone surpassing it at 4.9%, according to the latest available figures.

But this surge has not been even. Rather, it has favoured top-down, government-directed projects intended to boost competitiveness in fields such as artificial intelligence, robotics and materials, often in partnership with the private sector, consistent with the post-war focus on applied research that turned South Korea into a leader in semiconductor manufacturing and wireless communication networks.

New priorities, however, have taken root. For the past decade, the necessity to invest in basic research has been a common refrain from South Korean researchers and politicians alike. To maintain economic growth, the nation has sought to become a 'first mover', instead of simply a 'fast follower' — rhetoric embraced by the current liberal president, Moon Jae-in, and his predecessor, the conservative Park Geun-hye. The drive to compete more effectively in strategic fields continues, but South Korea has begun emphasizing bottom-up funding for small teams of researchers, betting that strength will flow from the creativity of its scientists.

Spending surge

On paper, the rhetoric is backed by funding. Under President Moon, who was first elected in 2017, the basic-research budget for South Korea's main funding agency, the National Research Foundation (NRF), has skyrocketed under a five-year plan to double it by 2022 to 2.5 trillion won (US$2 billion). The nation's budget for 2020 allows for a remarkable 18% increase from the previous year in government spending on research and development (R&D), to 24.2 trillion won.

In the private sector, which makes up roughly three-quarters of R&D spending, basic-research spending is also surging, including outlays by major conglomerates such as Samsung and LG Electronics. "Industry is asking universities to produce more and more PhDs," says Han Woong Yeom, a physicist at Pohang University of Science and Technology and vice-chair of Moon's presidential advisory council on science and technology. "The situation cannot be better than this, in terms of investment."

Spending surged over the past two decades on large national projects, such as a lunar orbiter scheduled for launch in 2022, a nanomaterials research effort dubbed the 'Creative Materials Discovery Program', and Korea's branch of the International Thermonuclear Experimental Reactor, the main experiment of which is under construction in southern France. "Previously, more emphasis was on the mission-oriented research towards immediate economic output," says microbiologist, Jung-Hye Roe, whom Moon tapped to lead the NRF in 2018.

Shift in emphasis

There is a new emphasis on bottom-up funding through NRF grants. Currently, the NRF accounts for slightly more than one-quarter of total government-funded research, with a budget of close to 7 trillion won in 2020. The amount spent on basic-research grants to principal investigators is now roughly equal to that spent on top-down, mission-oriented projects, and Roe hopes to boost the basic-research grants still further.

Recent highlights from these grants include: a gallium arsenide nanoresonator that can trap light and change its colour, work published in *Science*[1] in January and co-led by Hong-Gyu Park of Korea University in Seoul; atmospheric observations from Gosan-ri on Jeju Island as part of an international study[2], co-led by

© 2020 Springer Nature Limited. All rights reserved.



JONAS GRATZER/LIGHTROCKET VIA GETTY IMAGES

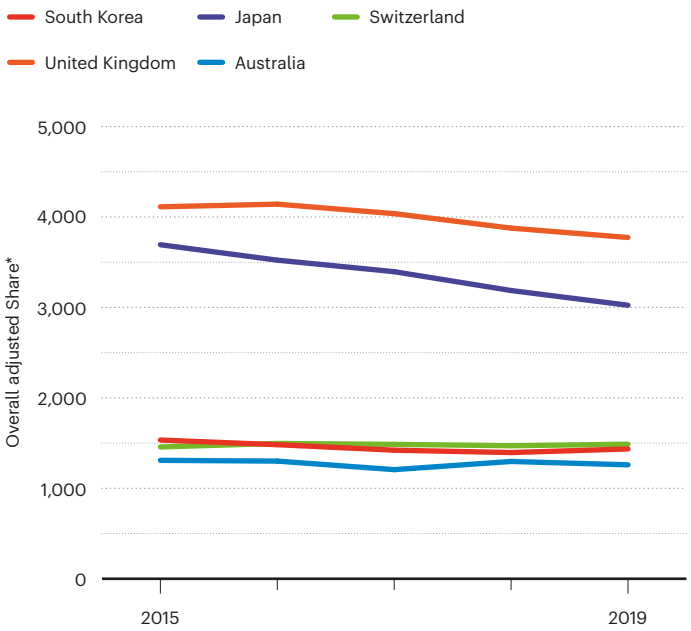
LG's Guide Robot at Incheon International Airport recognizes four languages.

© 2020 Springer Nature Limited. All rights reserved.

index

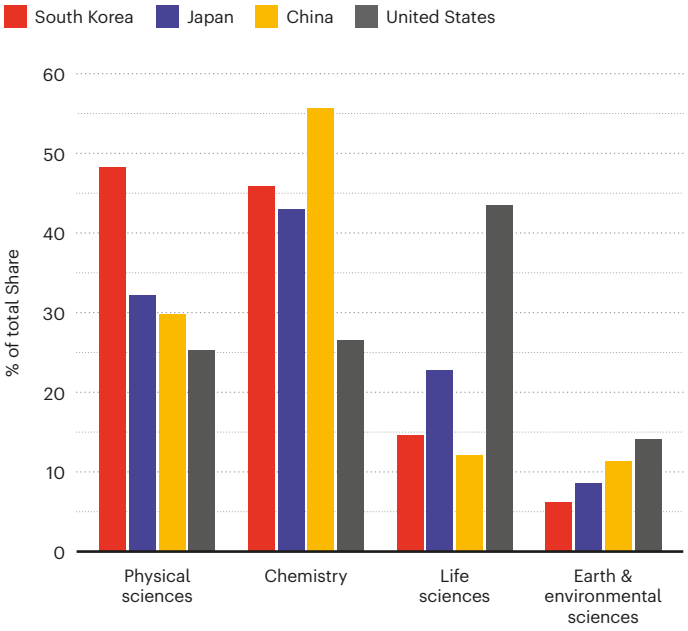
HOLDING STEADY—SELECTED COUNTRIES

In 2016 South Korea dropped from 8th to 9th among the top countries by Nature Index Share, and has retained 9th spot ever since, between Switzerland (8th in 2019) and Australia (10th). Japan is 5th; United Kingdom is 4th.



PHYSICAL STRENGTH

How much of each country's total Nature Index Share is taken up by each broad subject area. The physical sciences account for a greater proportion of Share in South Korea than any other subject.



SOURCE: NATURE INDEX

*When comparing data over time, Share values are adjusted to 2019 levels to account for the small annual variation in the total number of articles in the Nature Index journals.

Sunyoung Park of Kyungpook National University in Daegu and published in *Nature* in 2019, that identified eastern China as a major source of the emission of chlorofluorocarbons, chemicals that destroy the ozone layer that are banned under the Montreal Protocol; and the finding that the development of osteoarthritis is related to how the body processes cholesterol, in a 2019 *Nature* study[3], led by Je-Hwang Ryu of Chonnam National University in Gwangju and Jang-Soo Chun of the Gwangju Institute of Science and Technology.

The top-down division can claim successes too, including playing a key role in the nation's ability to quickly develop and produce diagnostic kits for COVID-19, says Roe. The first four companies to win urgent-use approval for their tests had all received NRF funding to research the production of such kits after the 2015 Middle East Respiratory Syndrome (MERS) outbreak in South Korea infected 186 people and killed 36. And in February, the Korea Aerospace Research Institute, responsible for the nation's space programme, launched the Cheollian 2B satellite; its main instrument, the Geostationary Environment Monitoring Spectrometer (GEMS) is designed to monitor air pollution in the Asia-Pacific region, making it the first of a new generation of satellites intended to study global air quality.

An emphasis on government-mandated metrics in hiring and promotion, which critics say incentivizes researchers to present work of little value in low-quality journals or conferences, remains to be addressed. Last May, the nation's education ministry reported that 574 researchers had participated in international conferences organized by 'predatory publishers'. "That happens because, so far, we have emphasized getting large numbers of quantitative, shallow publications," says Roe.

> **"This old-style culture may still be good enough to produce many papers, but not original ideas."**

"I think this old-style culture is prohibiting the booming of our creativity," says Yeom. "This may still be good enough to produce many papers, but not original ideas."

The outlook is uncertain. Not only does the expected heavy economic burden of recovering from the COVID-19 pandemic threaten to curb the growth in R&D budgets, but South Korea's traditional industries — including chip-making, ship-making and nuclear power — face increased competition from China. Political tensions with Japan that flared in July 2019, fuelled by that nation's legacy of colonial rule, threatened to cut off supplies of key materials for semiconductor production, exposing how heavily the nation's technology industry relies on Japanese manufacturers.

There is talk in policy circles of a science and technology new deal — a boost in funding and programmes to combat the effects of a less globalized world in which South Korea's access to scientific supplies and talent is reduced, says So Young Kim, director of the Korea Policy Center for the Fourth Industrial Revolution, a think tank affiliated with KAIST in Daejeon. Yeom says researchers stand to benefit from a public appetite for societal change, which also translates into support for increased science funding. Yeom and other policymakers are pushing for greater investment in research topics that meet the practical needs of South Korean society, such as air pollution, and an ageing population. In response, the Moon administration has begun an initiative to trace and reduce fine particulate matter across northeast Asia, and another to fight dementia.

Some suspect it will be initiatives such as these that do most to move South Korea away from the label 'fast follower'. "The definition of top-level research doesn't need to be the highest impact factor or highest citation count," says Roe. Pursuing research directions that benefit South Korean society is "making our own path, a new path, rather than following others".

Mark Zastrow is a science writer based in Seoul.

1. Koshelev, K. *et al*. *Science* **367,** 288–292 (2020).
2. Rigby, M. *et al*. *Nature* **569**, 546–550 (2019).
3. Choi, W.-S. *et al*. *Nature* **566**, 254–258 (2019).

© 2020 Springer Nature Limited. All rights reserved.

National vision

The vast IBS network signals South Korea's ambitions, but research focus may be adjusted after centres face scheduled scrutiny. **By Mark Zastrow**

South Korea's highest-profile investment in basic research is in the Institute for Basic Science (IBS), a network of research centres that comprises the nation's answer to Germany's Max Planck Society and Japan's RIKEN.

Founded in 2012 after a 2007 promise by conservative, Lee Myung-bak, during his successful campaign for the country's presidency, the institute sought to attract elite researchers from at home and abroad by giving successful individual applicants, as centre directors, full autonomy over their centre's research and a budget of 10 billion won (US$8.2 billion) per year for ten years.

The centres cover areas including dark matter, nanomaterials, genome engineering and climate change. But, after criticism that they were taking money away from the greater scientific community, the goal of 50 centres has been reduced to 30 for now, with average budgets trimmed to about 6 billion won.

IBS president, Noh Do Young, warns that the cuts are "likely to inhibit IBS's efforts to fulfil its missions and visions", arguing that the budgets are not large, considering centres can employ dozens of researchers each.

Many IBS researchers say the level of funding has helped their research reach new heights. For instance, in March 2020, the Center for RNA Research at Seoul National University used cutting-edge sequencers purchased with IBS funds to become one of the first research groups in the world to sequence the transcriptome – the total product of all expressed genes – of the coronavirus that caused the COVID-19 pandemic[1]. The sequencers can directly identify the molecules in RNA sequences, as well as where they are modified by other molecules. Narry Kim, a biochemist and the centre's director, says the work required an interdisciplinary team of virologists, microbiologists and computational scientists, which was enabled by IBS's steady support. "Without IBS funding, I don't think it would have been possible," she says.

Among other prominent IBS results, the Center for Genome Engineering supplied the gene-editing tools in a blockbuster 2017 study[2] in *Nature* that claimed the CRISPR–Cas9 gene-editing system was used to correct a mutation that causes heart disease in viable human embryos by replacing the mutated copy from the sperm with the correct gene from the egg. The results have been hotly debated, as some scientists argue there could be alternative explanations.

In 2018, results reported in *Nature*[3] by the Center for Underground Physics, from its laboratory in a subterranean power plant in Yangyang County near the country's east coast, ascribed constraints to the theorized particles that could make up dark matter. And in 2019, a collaboration including the Center for Climate Physics in Busan proposed in *Nature*[4] that the first ancestors of modern humans came from a prehistoric wetland in modern-day Botswana, roughly 200,000 years ago. The work combined the centre's supercomputer climate models with a genetic analysis of populations in Namibia and South Africa.

> **"If I cannot build a top-class research centre in eight years, either I'm not the right person for the job, or it's not possible in that environment."**



ASAN FOUNDATION

Narry Kim, director of the Center for RNA Research at Seoul National University.

Political strife

IBS came under heavy scrutiny in 2019, when reports surfaced in South Korean media of misappropriated funds and nepotism in hiring, which had allegedly been uncovered by government audits, at several centres. Doochul Kim, a physicist who was president of IBS at the time, said these were mostly the result of administrative errors and that the audits were politically motivated by liberal lawmakers. In late 2019, Kim announced reforms to IBS's administrative structure, to be carried out under his successor, Noh, who started in November 2019.

The year 2020 marks another significant milestone for IBS: the first group of centres are coming up for review. Unlike at the Max Planck Society, centre directors at IBS are not awarded a lifetime appointment, and centres are subject to a make-or-break review after eight years to determine if they will be extended beyond their original ten-year remit. "This review session will shape the future of IBS," says Noh. He says centres may close as a result, although new centres may take their place. (Noh adds that the process may be delayed by the COVID-19 pandemic, if crucial on-site visits by international reviewers, scheduled for July, cannot go ahead.)

The model has its advocates. "I like that pressure," says Andreas Heinrich, who left IBM Almaden in San Jose, California, to lead the IBS Center for Quantum Nanoscience in Seoul. "If I cannot build a top-class research centre in eight years, either I'm not the right person for the job, or it's not possible in that environment."

Mark Zastrow is a science writer based in Seoul.

1. Kim, D. *et al. Cell* **181**, 914–921 (2020).
2. Ma, H. *et al. Nature* **548**, 413–419 (2017).
3. Adhikari, G. *et al. Nature* **564**, 83–86 (2018).
4. Chan, E. K. F. *et al. Nature* **575**, 185–189 (2019).

© 2020 Springer Nature Limited. All rights reserved.



ED JONES/AFP VIA GETTY IMAGES

South Korea is a global leader in information and communication technologies.

A top-down reinvention

A concerted government push to make South Korea an innovation leader, backed by strong investment and systemic reform, has brought rapid and long-lasting results. **By Leigh Dayton**

South Korea's position as one of the world's most innovative nations is a remarkable achievement considering that, for the first half of the twentieth century, it was an agrarian-based Japanese colony, then a battle ground.

It is second only to Germany in Bloomberg's 2020 Innovation Index, having reigned at the top of the 60-country list for the previous 5 years. In the separate 2019 Global Innovation Index, published by Cornell University, INSEAD and the World Intellectual Property Organization, South Korea is at number 11 and Germany is in 9th place among the 129 countries ranked.

Both indices highlight South Korea's outstanding performance in research and development (R&D) intensity, an indicator based on R&D investment by government and industry and the number of researchers working in and between both sectors. For example, South Korea had the greatest share of researchers who moved from industry to academia in 2017 to 2019 among 71 countries, data from academic recruitment firm, League of Scholars, show.

Top-down success

The high R&D intensity that helped South Korea become a global leader in information and communication technologies has emerged from a historically 'top-down' innovation system that promotes "close collaboration between government, industry, and the academic community in the process of nation building", says Tim Mazzarol from the University of Western Australia in Perth, who specializes in innovation and entrepreneurship.

President Park Chung-hee drove South Korea's economic development between 1961, when he took power in a military coup, until 1979, when he was assassinated. Park shifted the economy from its post-war dependence on technology imports and the construction

© 2020 Springer Nature Limited. All rights reserved.

SAMSUNG'S TOP TEN

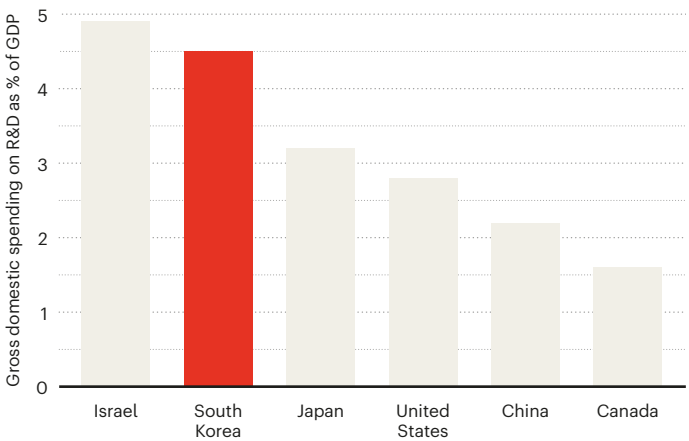
Samsung Group's top ten collaborating academic partners on articles in the Nature Index journals are split between United States and domestic institutions. Here they are ranked by bilateral collaboration score (CS), 2015–19. CS is derived by summing each institution's Share on the papers to which authors from both have contributed.

Rank	Institution	Country	Bilateral CS	Count*
1	Sungkyunkwan University	South Korea	75.07	159
2	Seoul National University	South Korea	21.10	41
3	Korea Advanced Institute of Science and Technology	South Korea	20.16	35
4	Stanford University	United States	19.29	31
5	University of California, Berkeley	United States	17.16	51
6	Korea University	South Korea	13.62	27
7	Yonsei University	South Korea	11.07	22
8	Harvard University	United States	9.67	26
9	Pohang University of Science and Technology	South Korea	8.82	16
10	California Institute of Technology	United States	8.35	12

*Count = Article count

SECOND TO ONE

Among all countries, Israel, at 4.9%, is the only one that spends a greater proportion of its Gross Domestic Product (GDP) on R&D than South Korea, at 4.5%. But the disparities in R&D, even among the selected countries shown, are vast in terms of the total spent, in 2018 ranging from US$16.3 billion (Israel) to $526 billion (China) and $551.5 billion (United States). South Korea spent $95.4 billion on R&D in 2018.



SOURCES: NATURE INDEX; OECD GROSS DOMESTIC SPENDING ON R&D (INDICATOR) HTTPS://DOI.ORG/10.1787/D8B068B4-EN (ACCESSED 8 MAY 2020)

of industrial facilities by foreign companies to focus on home-grown labour-intensive industries, such as clothing and textiles. Crucially, strong support for R&D was central to his first Five-Year Economic Development Plan in 1962 and manifest in his establishment of the Korea Institute of Science and Technology (KIST) in 1966, and the Ministry of Science and Technology the following year.

These instruments supported the emergence of large industrial groups called chaebols, which were owned and controlled by South Korean individuals or families. The government pushed the chaebols to invest heavily in R&D while shielding them from competition. With increased R&D intensity that focused on applied knowledge, chaebols such as LG, Lotte and Samsung were driven towards new heavy industries, including petrochemicals, car manufacturing and shipbuilding, as well as consumer electronics.

Samsung – the classic chaebol

Samsung is a case in point. The company that started life as a grocery trader in 1938 is now South Korea's largest chaebol, operating in industries as diverse as electronics, insurance, construction and shipbuilding. In 2018, it produced roughly 15% of the nation's gross domestic product.

Its founder, Lee Byung Chul, with help from government protectionist policies, expanded into textiles after the Korean War, electronics in the 1960s, then heavy industries, aerospace and computing during the 1970s and early 1980s. By the 1990s and 2000s, Samsung was a world leader in tablets and mobiles, and in the design and manufacture of computer chips.

The company is South Korea's leading corporate institution in the Nature Index by far, based on contributions to research articles published in the 82 high-quality natural science journals tracked by the Index. With a Share of 10.36 in 2019, it ranked 28th among the country's institutions overall, eclipsing its nearest rival in the corporate ranks, LG, which had a Share of 1.99. Samsung also features in each of South Korea's nine leading corporate-academic collaborative pairs in the Nature Index.

The most productive pairing is with Sungkyunkwan University (SKKU) in Seoul, with 159 joint articles between 2015 and 2019. Their collaboration is particularly strong in electrochemistry and the development of new energy sources such as lithium-ion batteries (J. K. Shon *et al. Nature Commun.* **7**, 11049; 2016). Other partnerships include Seoul National University in Seoul, (41 articles) and the Korea Advanced Institute of Science and Technology (KAIST) in Daejeon (35 articles).

Investing in the future

Park's successors continued to promote research and innovation as the driver of national economic and social advance. Government and industry investment in R&D soared, and basic-research capabilities were expanded. By the mid-1980s and early 1990s the government's attention had shifted to high-tech industries such as semiconductor design and manufacture. For instance, in 1981 it founded KAIST, which remains a leading national research university (see 'Manipulating brains with smartphones').

Targeted nation-building programmes were also established. In 1995, for example, the government began a US$1.5-billion, ten-year plan to build up the national broadband infrastructure and provide public programmes about maximizing its use.

The 1997 Asian Financial Crisis prompted many chaebols to shift from the reliance on low-value added exports characteristic of a 'tiger' economy towards technology and knowledge-intensive products and services such as semiconductors, mobile phones and mobile applications.

Working with chaebols, the government began developing regional innovation centres such as Gyeonggi, an area of nearly 13 million people surrounding Seoul, which is now regarded as the nation's economic and innovative powerhouse.

The centre brought industry R&D and production infrastructure together with local and national universities and research facilities. For instance, the Gyeonggi-based Samsung Electronics, Samsung's flagship subsidiary, is collaborating with SKKU Chemistry to develop

© 2020 Springer Nature Limited. All rights reserved.

a semiconductor material that can reduce the amount of radiation exposure while taking medical X-ray images. By 2010, South Korea had 105 regional innovation centres and 18 techno-parks, as well as 7 federal programmes to strengthen the competitiveness of industrial cluster programmes.

Although government funding continued to promote R&D spending and programmes to boost translational development and scientific, engineering and managerial expertise, the weight of major investment in R&D shifted to the corporate sector in search of patents and profits. Private R&D spending accounted for nearly 80% of South Korea's total R&D spending in 2019, ahead of leading innovative nations such as Germany, Sweden and Switzerland, at 70%. The shift was supported by R&D tax incentives and importation of foreign technology.

The new breed

In the 2010s, small to medium-sized businesses in biotechnology, artificial intelligence and cybersecurity, and broadband-based firms began to emerge. Founded by a new generation of entrepreneurs, they were backed by government funding and supported by the national technological infrastructure.

Woowa Brothers is one example, among many, of the strategy's success. The Seoul-based 2010 start-up exploited the national broadband to build a mobile food-delivery application connecting restaurants, customers and riders.

In December 2018, Woowa joined the 'unicorn' club — a rare status denoting a privately held start-up valued at more than US$1 billion — with investment from national and international venture-capital sources. In December 2019, Germany's Delivery Hero bought the company in a $4-billion deal that will see co-founder and chief executive, Kim Bong Jin, manage the Asian business, including South Korea, Vietnam and Hong Kong. A delivery robot, self-driving technology and an online customer and revenue-management system for restaurants are in development.

The South Korean government's systematic approach has been the crucial factor in creating an innovative economy adept at turning ideas from laboratories into products and industries. Martin Hemmert, an expert in east Asian innovation systems, at Korea University, adds that the cultural mindset evident in South Korea helps. "Complacency is not on the cards. The glass is always half empty," he says.

Even so, as Mazzarol concludes: "It's a miracle when you consider where Korea was."

Leigh Dayton is science and innovation writer based in Sydney, Australia.

MANIPULATING BRAINS WITH SMARTPHONES

A team of researchers at KAIST — not to be confused with the KIST, with which it was initially integrated — has fulfilled a dream of neuroscientists worldwide.

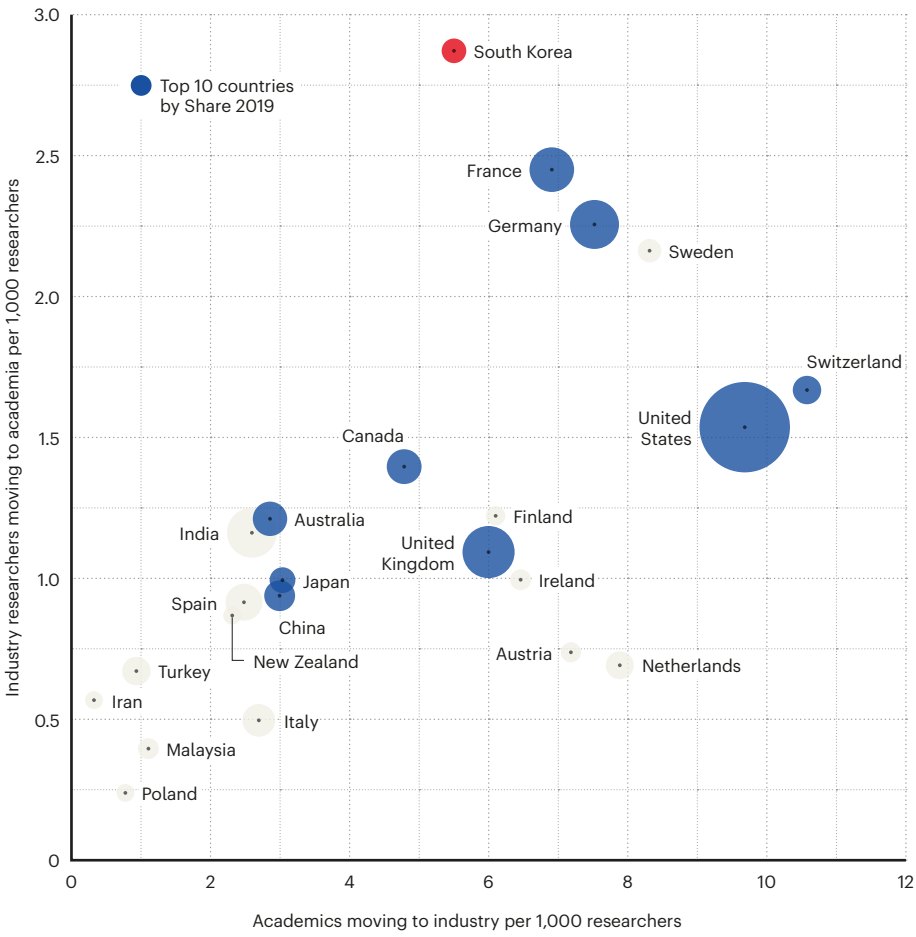
Working with colleagues at the University of Washington in Seattle, they have built a novel device capable of remotely controlling the brain circuitry of mice, via a smartphone. It is the first wireless neural device that can continuously deliver multiple drugs and coloured light beams to control brain circuits. Until now, researchers needed rigid metal tubes and optical fibres to accomplish the task.

The device could speed up the study of diseases such as Parkinson's, Alzheimer's, addiction, depression and pain, says team leader and electrical engineer, Jeong Jae-Woong. Weighing 2 grams, it uses LEGO-like replaceable drug cartridges, a probe the thickness of a human hair, and powerful, low-energy Bluetooth to deliver the drugs and light, which turn neurones on or off without hurting the rodents. This 'plug-and-play' interface was the major challenge, Jeong says.

After two years of laboratory and animal trials, the proof-of-concept paper was published in *Nature Biomedical Engineering* (R. Qazi *et al. Nature Biomed. Eng.* **3**, 655–669; 2019) by the team of neuroscientists and engineers from electrical, mechanical and software backgrounds. Jeong plans to commercialize the device and technology.

TALENT SWAP

Cross-sectoral moves between industry and academia per 1,000 researchers 2017–19 are shown for the top 10 countries by Share in the Nature Index. South Korea is a global outlier for its high proportion of researchers moving from industry to academia. For shifts from academia to industry, it is in the middle of the pack, with a slightly higher proportion than Canada and slightly less than the UK. Switzerland then the US have by far the greatest share of researchers moving from academia to industry. Bubbles are sized according to research scale, a function of the number of research institutions in each country covered by the data set.



SOURCE: LEAGUE OF SCHOLARS

© 2020 Springer Nature Limited. All rights reserved.

Upwardly mobile

With growing numbers of researchers coming and going, South Korea's scientific landscape is becoming more diverse, and more productive. **By Chris Woolston**

The proportion of academics who have relocated to South Korea from other countries in the past three years is higher than the global average, suggesting that the country's drive to end its comparative scientific isolation may be bearing fruit.

"It's becoming a hub for global talent," says Paul McCarthy, chief executive of League of Scholars, an academic recruitment firm based in Sydney, Australia. The firm is using publication records to track mobility and productivity of more than 2 million researchers from around the world, including 26,697 researchers in South Korea, more than 95% of whom work in science and technology.

Author affiliations in the League of Scholars data set show that 4.3–4.9% of all academics in South Korea had moved from abroad between 2017 and 2020, whether foreign-born researchers relocating from elsewhere, or those returning home after working abroad. That's higher than the global average of 3.7%. The data on these globally mobile researchers indicate that the country has recently attracted an influx of productive and successful scientists, according to McCarthy.

High-impact scientists

Globally mobile researchers in South Korea tend to thrive. Researchers who relocated to South Korea since 2017 have been found to be more than 50% more productive than other academics in the country, measured by the median number of publications they authored or co-authored with at least ten citations in the previous five years. They also have 39% more lifetime citations than researchers who remained in South Korea during the data-collection period. In another measure of performance, the median annualized *h*-index of globally mobile researchers in Korea is 21% higher than their colleagues who remained at home.

Researchers who tend to be internationally mobile are more productive wherever they are found, not just in South Korea, notes Cornelia Lawson, a science policy researcher at the University of Manchester, UK.

Lawson says that working abroad gives South Korean researchers important insights into what it takes to get published in journals with an international reach. South Korean researchers who publish largely or exclusively in domestically published journals tend to miss out on the exposure and citations that come with higher-impact journals, she says. "A number of countries in Asia have a strong local publication culture," she says. "Domestic scientists may not understand or feel the need to publish internationally."

When it comes to sheer number of publications, however, home-grown researchers in South Korea can compare favourably with, or outperform, their colleagues from elsewhere. A 2014 study[1] published in the journal *Minerva* found that researchers with domestic degrees in South Korea, Hong Kong and Malaysia published more refereed papers and books or book chapters in the natural sciences, engineering and biomedical sciences than researchers holding foreign degrees. Many of those papers were in domestic journals, where native researchers would have a clear publication advantage.

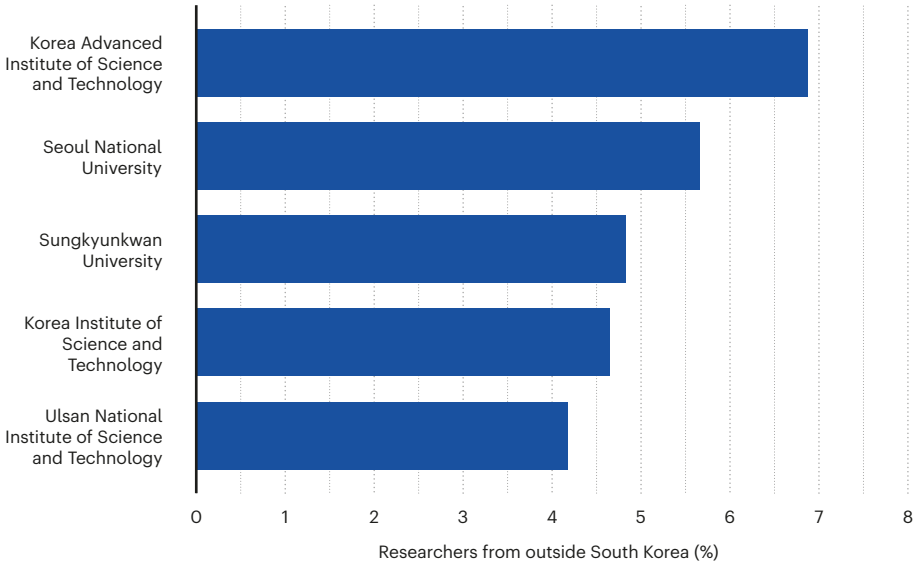
For Moon Kee Choi, a bioengineer who works on wearable and implantable electronic devices at the Ulsan National Institute of Science and Technology, the monthly 'happy hour' at the University of California, Berkeley (UC Berkeley), where she was a postdoctoral researcher from 2017 to 2019, was a career boon. Bioengineers exchanged ideas while buying snacks and drinks. "Many people in South Korea collaborate with others, but until now, that collaboration is largely domestic," she says. "Doing an international postdoc is a good way to make connections."

'Boomerang' researchers such as Choi, who return home to South Korea, often have experience with both domestic and international publishing, and benefit from broader networks and more opportunities for international collaboration, she says.

For Choi, those collaborations have translated into publications. In 2019, she co-authored

GLOBAL TALENT MAGNETS

South Korea has a higher-than-average proportion of academics who have relocated from an overseas institution in the last 3 years. KAIST has attracted the biggest share of this valuable globally mobile group: the so-called border-crossers have a 21% higher impact (based on annualized *h*-index) and 50% higher output (based on articles over the past 5 years with at least 10 citations) than their stay-at-home colleagues.

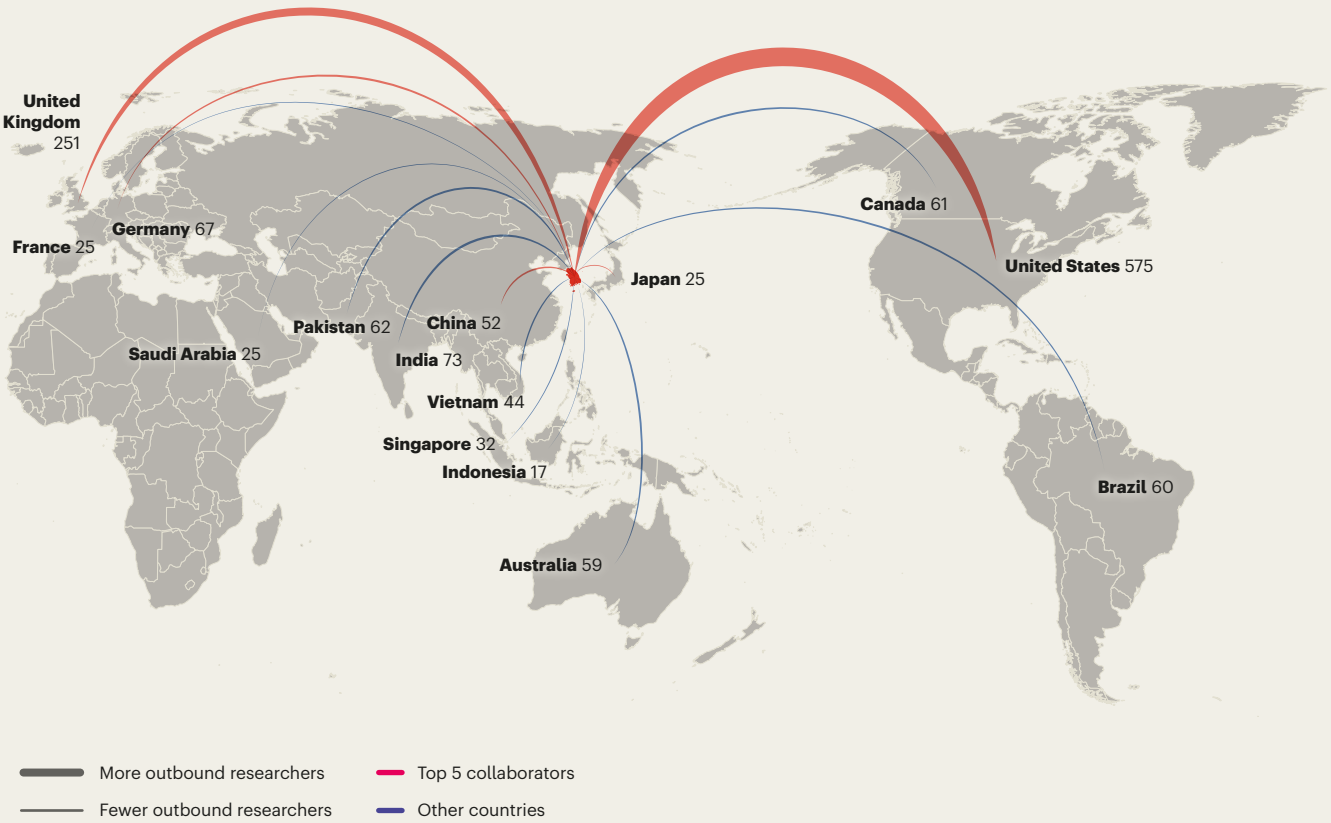


SOURCE: LEAGUE OF SCHOLARS

© 2020 Springer Nature Limited. All rights reserved.

KNOWLEDGE TRANSFERS

Destinations of researchers who have relocated from South Korea since 2017. The US, which is South Korea's top collaborating country on articles in the Nature Index journals, is also the favoured location of its research diaspora, landing 575 recruits in the past 3 years. China and Japan are respectively 2nd and 3rd among South Korea's research collaborators, yet have received comparatively few of its researchers over the period.



SOURCE: LEAGUE OF SCHOLARS

a paper[2] on molybdenum sulfide capsules for cellular microscopy with three UC Berkeley researchers. Sanghyuk Wooh, a materials scientist who studies surfaces at Chung-Ang University in Seoul, suggests that intense competition that rewards higher achievers is another reason why Korean 'boomerang' researchers may be especially productive and impactful. Wooh completed a postdoc at the Max Planck Institute for Polymer Research in Mainz, Germany, before returning to South Korea in 2017. South Korean PhD graduates going on to postdoctoral positions in foreign countries tread a well-travelled path, but the return trip is more challenging, he says. "Most of the researchers who go overseas want to come back to South Korea," he says. "But there are a limited number of faculty positions, so it's very competitive."

Reaching out

The path may be somewhat easier for foreigners. The government has made a concerted push to attract and retain international talent in recent years, including several special initiatives to hire overseas researchers. One example is the World Class University Project, a programme that started targeting high-profile international researchers in 2008.

International researchers hired through such schemes often have a relatively easy path to promotion, says Stephanie Kim, a higher-education researcher at Georgetown University in Washington DC, who authored a 2016 study[3] of Western academics at South Korean universities. They also enjoy fringe benefits, including, in some cases, free housing. On the downside, they are not always well integrated into the rest of the university. "They have some advantages, but they are also marginalized and sequestered," says Kim.

Still, she adds, the resources and support can make South Korea an especially rewarding place to do research. "Unlike the United States, where funding for higher education is always dwindling, there's actually money being poured into the higher-education sector in South Korea," says Kim.

Although the League of Scholars data suggest that South Korea is making headway in an attempt to internationalize, Lawson notes that the country still lags behind other countries in terms of the diversity of its research faculty. According to a 2019 report by the Organisation for Economic Co-operation and Development, only 6.3% of research faculty at South Korean universities in 2017 were foreign-born, a slight drop from previous years. By comparison, 28% of full-time science and engineering faculty in the United States were born elsewhere, according to a 2018 report from the US National Science Foundation.

Chris Woolston is a freelance writer in Billings, Montana.

1. Shin, J. C., Jung, J., Postiglione, G. A. & Azman, N. *Minerva* **52**, 467–487 (2014).
2. Yang, J., Choi, M. K., Ercius, P. & Zheng, H. *Microsc. Microanal.* **25** (Suppl. 2), 1406–1407 (2019).
3. Kim, S. K. *Comp. Educ.* **52**, 78–90 (2016).

© 2020 Springer Nature Limited. All rights reserved.